ONLINE CLUSTERING OF BANDITS WITHOUT FEATURE DIVERSITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Online clustering of bandits shares information across heterogeneous users by pooling observations within groups of users whose similarity is inferred from observed rewards. Existing methods typically group users when their preference vectors have the same values in all coordinates. To accumulate enough information to infer such groups, these methods rely on either distributional assumptions on feature generation or access to an action set spanning the entire d-dimensional feature space; neither requirement is guaranteed to hold in practice. To dispense with such requirements, we group users when their expected rewards are the same for every available action, even if their preference vectors are different. We propose `TRIBE`, which explores a basis of the space spanned by the available actions, groups users through a novel and computationally efficient statistical test, and selects actions using the observations pooled within each inferred group. Over T rounds, our algorithm achieves a regret bound with the leading term $\widetilde{\mathcal{O}}(d\sqrt{CT})$, where C is the number of clusters, matching the rate attainable when the cluster structure is known a priori. Numerical experiments on synthetic and real-world datasets support our theoretical findings and confirm that `TRIBE` outperforms the baselines.

1 INTRODUCTION

Linear stochastic bandits (Auer, 2002; Abbasi-yadkori et al., 2011; Agrawal & Goyal, 2013) model the expected reward of an action as the inner product between the action's feature vector and an unknown parameter vector, representing a preference vector shared across all interactions. Such a shared model treats all users as homogeneous, whereas users of personalized services such as recommender systems (Li et al., 2010) and mobile health (Tomkins et al., 2021) respond differently to the same action. Learning an independent bandit model for each user accommodates heterogeneity, but each model is then estimated only from the observations of one user, requiring many samples for effective learning. Moreover, such complete separation is often unnecessarily conservative in practice, since users are rarely entirely distinct (Li et al., 2016). Online clustering of bandits bridges these extremes by learning user groups and sharing observations within each group (Gentile et al., 2014).

Prior studies typically posit an unknown cluster structure where users are grouped when their preference vectors have the same values in all coordinates. We write C_{lat} for the number of such clusters, i.e., the number of distinct preference vectors. To recover this parameter-defined cluster structure from data, one line of work imposes stochastic feature diversity, under which feature vectors are generated IID from a distribution with a positive-definite second-moment matrix (Gentile et al., 2014; Li & Zhang, 2018; Li et al., 2019; Liu et al., 2022; Wang et al., 2023a;b; Yang et al., 2024; Li et al., 2025). Alternatively, Cherkaoui et al. (2025) assume access to a fixed action set spanning the entire feature space $\mathbb{R}^d$. Both requirements make every direction of $\mathbb{R}^d$ available for exploration, either through feature generation itself or through a fixed action set spanning $\mathbb{R}^d$, so that each user's complete preference vector can be identified. Such distributional assumptions can be restrictive in practice (Li et al., 2025), and fewer than d actions cannot span $\mathbb{R}^d$.

More fundamentally, such full identifiability is a stronger condition than required for decision-making problems: expected rewards, and hence which actions are optimal, depend only on the parameter's projection onto the span of the action set (Lattimore & Szepesvari, 2017; Yang & Tan, 2022). This discrepancy motivates our central research question: can we learn and exploit user clusters using only the reward-relevant information available through the action set, without identifying each user's

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

complete preference vector? The arm-dependent grouping of Gentile et al. (2017) also compares users through expected rewards, but their analysis still relies on stochastic feature diversity. We therefore focus on *observable clusters* (see Definition 1), which group users whose expected rewards are the same for every available action, and recover these clusters without requiring stochastic feature diversity or an action set spanning $\mathbb{R}^d$. Our main contributions are summarized as follows.

- **Algorithmic design for cluster recovery.** We propose TRIBE, which identifies clusters by combining (i) round-robin exploration over a basis of the available actions and (ii) a novel harmonic-mean test that refines the cluster estimates. For action selection, TRIBE adopts a UCB strategy that pools all observations from the served user's current neighborhood. The resulting design recovers the observable clusters without requiring stochastic feature diversity or an action set spanning $\mathbb{R}^d$ (see Section 4).
- **Sample-efficient learning.** TRIBE achieves a regret bound with the leading term $\widetilde{\mathcal{O}}(d\sqrt{CT})$ over T rounds, where C is the number of observable clusters. The leading term matches the bound obtained by running OFUL (Abbasi-yadkori et al., 2011) within each observable cluster when the cluster memberships are known. Since $C \leq C_{\mathrm{lat}}$ (see Lemma 3), the leading term is no larger than $\widetilde{\mathcal{O}}(d\sqrt{C_{\mathrm{lat}}T})$, the rate of existing guarantees (Gentile et al., 2014; Li et al., 2019; 2025) (see Section 5.1). The analysis further extends to IID time-varying action sets without a positive-definite second-moment matrix (see Appendix D).
- **Reduced computational complexity.** Our forced exploration strategy enables all users' design matrices, which store exploration samples, to share a common structure (see Lemma 1) and hence a common eigenbasis. This strategy reduces each pairwise separation test to a coordinate-wise computation in $\mathcal{O}(B)$ time, where $B \leq d$ denotes the number of basis vectors in the action set. In contrast, the ellipsoid-intersection test of Cherkaoui et al. (2025) requires a pairwise eigendecomposition and an iterative numerical minimization costing $\mathcal{O}(d^3 + \tau d)$, where τ is the number of minimization iterations (see Section 5.2).
- **Empirical validation.** Experiments on both synthetic and real-world datasets with varying cluster separation demonstrate that TRIBE consistently achieves lower cumulative regret than the baselines while maintaining competitive runtime scaling (see Section 6).

2 Related Work

To obtain a regret bound that depends on the number of clusters sublinearly, earlier analyses in the online clustering of bandits recover parameter-defined clusters, under stochastic feature diversity and additional variance restrictions (Gentile et al., 2014; Li & Zhang, 2018; Li et al., 2019; Liu et al., 2022). Li et al. (2025) relax the variance restrictions through forced exploration while retaining stochastic feature diversity. Cherkaoui et al. (2025) dispense with stochastic feature diversity but require an arm set spanning $\mathbb{R}^d$, and their regret bound depends on the minimum cluster proportion rather than the number of clusters. We instead define observable clusters (see Definition 1), which group users whose expected rewards are the same for every available arm, and learn these clusters with neither assumption.

Gentile et al. (2017) also group users by expected rewards rather than by equality of preference vectors, but define a separate user neighborhood for each arm. To exploit the expected number of clusters in their regret bound, stochastic feature diversity and additional variance restrictions are required again; without such assumptions, a data-dependent term in their bound can grow linearly in T. Their algorithm additionally requires a separation parameter as input, which serves as a lower bound on the expected-reward gap between users in different neighborhoods of the same arm and determines when observations are shared among neighbors. In contrast, TRIBE attains a leading regret term of $\widetilde{\mathcal{O}}(d\sqrt{CT})$ without stochastic feature diversity or a separation parameter.

We adopt graph-based cluster identification with direct-neighborhood pooling. Graph-based methods refine cluster estimates by removing edges from a complete user graph through a statistical test for separation (Gentile et al., 2014; Li & Zhang, 2018; Liu et al., 2022; Wang et al., 2023a;b; Li et al., 2025; Cherkaoui et al., 2025). Recent approaches pool observations over direct neighbors instead of connected components, so that a cross-cluster edge does not bias an entire component (Wang et al., 2023a; Li et al., 2025; Cherkaoui et al., 2025). Most of the statistical tests compare estimated preference vectors in the isotropic ℓ_2 norm (Gentile et al., 2014; Li et al., 2019; 2025), which requires



(a) Latent clusters (3 clusters) (b) Observable clusters (2 clusters)

Figure 1: Latent clusters versus observable clusters in $\mathbb{R}^2$, where the arm set spans a line W. (a) Users (filled dots) sharing each of three distinct latent parameters (hollow diamonds) form three latent clusters, thus $C_{\text{lat}} = 3$. (b) Two of the three latent parameters have the same projection onto W (filled diamonds), so the users form two observable clusters, i.e., $C = 2$.

stochastic feature diversity to make the estimates accurate in every direction. Cherkaoui et al. (2025), on the other hand, test the exact intersection of confidence ellipsoids through numerical minimization costing $\mathcal{O}(d^3 + \tau d)$ per pair. Our harmonic-mean test also exploits the ellipsoidal geometry but costs $\mathcal{O}(d)$ per pair. A more comprehensive literature review is presented in Appendix A.

3 Preliminaries

3.1 Notation

We write scalars in non-bold letters, vectors in bold lowercase, and matrices in bold uppercase. For $n \in \mathbb{N}$, $[n] \coloneqq \{1, \dots, n\}$, and $|S|$ denotes the cardinality of a set S. For a vector $\boldsymbol{v} \in \mathbb{R}^d$, we write $\|\boldsymbol{v}\| \coloneqq \sqrt{\boldsymbol{v}^\top \boldsymbol{v}}$ for its ℓ_2-norm. For a positive-semidefinite matrix $\boldsymbol{D} \in \mathbb{R}^{d \times d}$, we define the weighted seminorm $\|\boldsymbol{v}\|_{\boldsymbol{D}} \coloneqq \sqrt{\boldsymbol{v}^\top \boldsymbol{D} \boldsymbol{v}}$; when $\boldsymbol{D}$ is positive definite, $\|\cdot\|_{\boldsymbol{D}}$ is a norm. For symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same dimension, we write $\boldsymbol{A} \succeq \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite, and $\boldsymbol{A} \succ \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is positive definite. We write $\boldsymbol{I}_d$ for the $d \times d$ identity matrix, $\boldsymbol{0}_{p \times q}$ for the $p \times q$ zero matrix, $\boldsymbol{0}_d$ for the d-dimensional zero vector, and $\mathbf{1}\{E\}$ for the indicator of an event E.

3.2 Problem Formulation

Consider a population of N users, and each user $i \in [N]$ is associated with an unknown preference vector $\boldsymbol{\theta}_i^\star \in \mathbb{R}^d$. The learner is given an arm set $\mathcal{A} \coloneqq \{\boldsymbol{x}_k \in \mathbb{R}^d : k \in [K]\}$, and every arm is available at every round; Appendix D extends the analysis to IID time-varying arm sets. Learning proceeds sequentially over a finite horizon T. At each round $t \in [T]$, a user u_t is drawn independently from an unknown distribution over $[N]$, allowing non-uniform arrival probabilities: for all $i \in [N]$,

$$\mathbb{P}(u_t = i) = \pi_i > 0, \tag{1}$$

where $\sum_{i=1}^N \pi_i = 1$. We write $\pi_{\min} \coloneqq \min_{i \in [N]} \pi_i$, so that $N\pi_{\min} \in (0, 1]$. Our arrival model recovers the IID uniform user-arrival model (Gentile et al., 2014; Li & Zhang, 2018; Wang et al., 2023a; Li et al., 2025) by setting $\pi_i = 1/N$ for all $i \in [N]$, in which case $N\pi_{\min} = 1$.

After observing u_t, the learner selects an arm index $a_t \in [K]$ and observes the reward

$$y_t = \boldsymbol{x}_{a_t}^\top \boldsymbol{\theta}_{u_t}^\star + \epsilon_t, \tag{2}$$

where ϵ_t denotes the reward noise. We impose the following two assumptions.

Assumption 1 (Sub-Gaussian noise). *For every $t \in [T]$, let $\mathcal{F}_{t-1} \coloneqq \sigma(\mathcal{A}, (u_s, a_s, y_s)_{s<t}, u_t)$ denote the information available before selecting a_t. The noise ϵ_t is conditionally 1-sub-Gaussian given $\mathcal{F}_{t-1}$; that is, for every $\eta \in \mathbb{R}$,*

$$\mathbb{E}\left[\exp(\eta \epsilon_t) \mid \mathcal{F}_{t-1}\right] \le \exp\left(\eta^2/2\right).$$

Assumption 2 (Bounded norms). *For all $k \in [K]$ and $i \in [N]$, $\|\boldsymbol{x}_k\| \le 1$ and $\|\boldsymbol{\theta}_i^\star\| \le 1$.*

Assumption 1 implies $\mathbb{E}\left[\epsilon_t \mid \mathcal{F}_{t-1}\right] = 0$, and hence the expected reward is $\mathbb{E}\left[y_t \mid \mathcal{F}_{t-1}\right] = \boldsymbol{x}_{a_t}^\top \boldsymbol{\theta}_{u_t}^\star$, while Assumption 2 fixes the scale of the expected rewards.

Following Gentile et al. (2014), users whose preference vectors, called *latent parameters*, have the same values in all coordinates constitute a *latent cluster*, and we denote the number of latent clusters by $C_{\mathrm{lat}} := |\{\boldsymbol{\theta}_i^\star : i \in [N]\}| \ll N$. The latent clusters form a partition of $[N]$; throughout the paper, we use "cluster" for an individual group of users and "partition" for the collection of all such groups. The expected rewards of the available arms, however, depend on a preference vector only through the directions spanned by the arms. Write $W := \mathrm{span}(\mathcal{A}) \subseteq \mathbb{R}^d$, and let Π_W and $\Pi_{W^\perp}$ denote the orthogonal projections onto W and onto the orthogonal complement $W^\perp$. Every preference vector decomposes as $\boldsymbol{\theta}_i^\star = \Pi_W \boldsymbol{\theta}_i^\star + \Pi_{W^\perp} \boldsymbol{\theta}_i^\star$, and every arm lies in W, so the second component is orthogonal to every arm. Consequently, for all $i \in [N]$ and $k \in [K]$,

$$\boldsymbol{x}_k^\top \boldsymbol{\theta}_i^\star = \boldsymbol{x}_k^\top \Pi_W \boldsymbol{\theta}_i^\star. \tag{3}$$

We therefore refer to $\Pi_W \boldsymbol{\theta}_i^\star$ as the *observable parameter* of user i and group users as follows.

Definition 1 (Observable clusters)**.** For any users $i, j \in [N]$, define the relation $\sim_{\mathrm{obs}}$ by

$$\begin{aligned} i \sim_{\mathrm{obs}} j &\iff \boldsymbol{x}_k^\top \boldsymbol{\theta}_i^\star = \boldsymbol{x}_k^\top \boldsymbol{\theta}_j^\star \quad \forall\, k \in [K] \\ &\iff \Pi_W \boldsymbol{\theta}_i^\star = \Pi_W \boldsymbol{\theta}_j^\star. \end{aligned}$$

The equivalence classes induced by $\sim_{\mathrm{obs}}$ are called *observable clusters*. Let C denote the number of observable clusters, and let $m : [N] \to [C]$ denote an unknown surjective membership map such that $m(i) = m(j)$ if and only if $i \sim_{\mathrm{obs}} j$. For each $c \in [C]$, let $\mathcal{I}_c := \{i \in [N] : m(i) = c\}$ denote the c-th observable cluster and $\boldsymbol{\vartheta}_c^\star \in W$ the observable parameter shared by the users in $\mathcal{I}_c$, so that $\Pi_W \boldsymbol{\theta}_i^\star = \boldsymbol{\vartheta}_c^\star$ for every $i \in \mathcal{I}_c$; the collection $\{\mathcal{I}_c\}_{c \in [C]}$ is called the *observable partition*.

As illustrated in Figure 1, each observable cluster is a union of latent clusters, so $C \le C_{\mathrm{lat}}$ (see Lemma 3), and the latent and observable partitions coincide if and only if Π_W maps distinct latent parameters to distinct observable parameters, which holds in particular when $W = \mathbb{R}^d$. By Definition 1, the reward model in Eq. (2) can equivalently be written as

$$y_t = \boldsymbol{x}_{a_t}^\top \boldsymbol{\vartheta}_{m(u_t)}^\star + \epsilon_t,$$

so the expected rewards of the available arms are determined by C distinct observable parameters rather than by all C_{lat} distinct preference vectors. Moreover, since $\Pi_W \boldsymbol{\theta}_i^\star$ and $\Pi_{W^\perp} \boldsymbol{\theta}_i^\star$ are orthogonal, the Pythagorean theorem and Assumption 2 give, for every $c \in [C]$ and $i \in \mathcal{I}_c$,

$$\|\boldsymbol{\vartheta}_c^\star\|^2 = \|\Pi_W \boldsymbol{\theta}_i^\star\|^2 = \|\boldsymbol{\theta}_i^\star\|^2 - \|\Pi_{W^\perp} \boldsymbol{\theta}_i^\star\|^2 \le \|\boldsymbol{\theta}_i^\star\|^2 \le 1.$$

Let $a_t^\star \in \mathrm{argmax}_{k \in [K]} \boldsymbol{x}_k^\top \boldsymbol{\vartheta}_{m(u_t)}^\star$ denote an optimal arm index at round t. The learner aims to minimize the cumulative regret $\mathrm{Reg}(T) := \sum_{t=1}^T (\boldsymbol{x}_{a_t^\star}^\top \boldsymbol{\vartheta}_{m(u_t)}^\star - \boldsymbol{x}_{a_t}^\top \boldsymbol{\vartheta}_{m(u_t)}^\star)$. By Definition 1, $\mathrm{Reg}(T)$ equals the regret defined with $\boldsymbol{\theta}_{u_t}^\star$.

Finally, the difficulty of recovering the observable clusters depends on how far apart distinct observable parameters are along the available arms. If $W = \{\mathbf{0}_d\}$, every arm has zero expected reward and $\mathrm{Reg}(T) = 0$. Therefore, we focus on the case where $W \neq \{\mathbf{0}_d\}$ throughout the paper, and let $\mathcal{B} \subseteq \mathcal{A}$ be a basis of W. We write $B := |\mathcal{B}| = \dim(W) \le d$ and $\boldsymbol{\Sigma}_{\mathcal{B}} := B^{-1} \sum_{\boldsymbol{x} \in \mathcal{B}} \boldsymbol{x}\boldsymbol{x}^\top$.

Definition 2 (Minimum observable cluster gap)**.** For $C \ge 2$, the minimum observable cluster gap is

$$\Delta_{\mathcal{B}} := \min_{c \neq c'} \|\boldsymbol{\vartheta}_c^\star - \boldsymbol{\vartheta}_{c'}^\star\|_{\boldsymbol{\Sigma}_{\mathcal{B}}} = \min_{c \neq c'} \sqrt{\frac{1}{B} \sum_{\boldsymbol{x} \in \mathcal{B}} (\boldsymbol{x}^\top \boldsymbol{\vartheta}_c^\star - \boldsymbol{x}^\top \boldsymbol{\vartheta}_{c'}^\star)^2}.$$

Thus, $\Delta_{\mathcal{B}}$ is the root-mean-square difference in expected rewards, over the arms in $\mathcal{B}$, between the two closest observable clusters. Unlike the isotropic ℓ_2 parameter gaps used in prior analyses (Gentile et al., 2014; Li & Zhang, 2018; Li et al., 2019; 2025), $\Delta_{\mathcal{B}}$ measures differences only through the reward-relevant directions exposed by $\mathcal{B}$, as in the design-weighted gaps of Cherkaoui et al. (2025). Since $\mathcal{B}$ spans W, $\boldsymbol{\Sigma}_{\mathcal{B}}$ is positive definite when restricted to W, although $\boldsymbol{\Sigma}_{\mathcal{B}}$ may be singular on $\mathbb{R}^d$. Distinct observable parameters differ by a nonzero vector in W, so each pairwise $\boldsymbol{\Sigma}_{\mathcal{B}}$-distance is strictly positive, and $\Delta_{\mathcal{B}} > 0$ follows from the finiteness of C without a separate separation assumption. Moreover, for every $\boldsymbol{x} \in \mathcal{B}$ and $c \neq c'$, the Cauchy-Schwarz and triangle inequalities, Assumption 2, and $\|\boldsymbol{\vartheta}_c^\star\| \le 1$ give

$$\left|\boldsymbol{x}^\top (\boldsymbol{\vartheta}_c^\star - \boldsymbol{\vartheta}_{c'}^\star)\right| \le \|\boldsymbol{x}\| \|\boldsymbol{\vartheta}_c^\star - \boldsymbol{\vartheta}_{c'}^\star\| \le \|\boldsymbol{\vartheta}_c^\star\| + \|\boldsymbol{\vartheta}_{c'}^\star\| \le 2,$$

so $0 < \Delta_{\mathcal{B}} \le 2$. Note that $\Delta_{\mathcal{B}}$ is unknown to the learner, and we impose no lower bound on $\Delta_{\mathcal{B}}$.

Algorithm 1 TRIBE (Testing Reward equivalence In the action span via Basis Exploration)

Input: $\mathcal{A} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K\}$, regularization parameter $\lambda \geq 1$, confidence parameter $\delta \in (0,1)$, phase parameter $p \in (1/2, 1)$, UCB coefficient β, phase-scale constant C_0
Choose an ordered spanning basis $\mathcal{B} = \{\boldsymbol{x}_{b_1}, \ldots, \boldsymbol{x}_{b_B}\}$
$B \leftarrow |\mathcal{B}|, \quad \boldsymbol{\Sigma}_{\mathcal{B}} \leftarrow B^{-1} \sum_{h=1}^{B} \boldsymbol{x}_{b_h} \boldsymbol{x}_{b_h}^\top$
Initialize the complete graph $\mathcal{G}_{-1} = ([N], E(\mathcal{G}_{-1}))$
For all $i \in [N]$,
    set main statistics $\boldsymbol{V}_{i,0} \leftarrow \lambda \boldsymbol{I}_d$, $\boldsymbol{b}_{i,0} \leftarrow \boldsymbol{0}_d$
    set exploration statistics $\boldsymbol{V}_{i,\mathrm{e},0} \leftarrow \lambda \boldsymbol{I}_d$, $\boldsymbol{b}_{i,\mathrm{e},0} \leftarrow \boldsymbol{0}_d$, $n_{i,\mathrm{e},0} \leftarrow 0$
    set pointer $p_i \leftarrow 1$ and buffer $\mathcal{D}_i \leftarrow \varnothing$
$t \leftarrow 0$
for $s = 0, 1, 2, \ldots$ **do**
    $M_s \leftarrow C_0 2^s, \quad L_s \leftarrow \lceil 2^{ps/(1-p)} M_s \rceil$
    for $\ell = 1, \ldots, M_s + L_s$ **(terminate the algorithm if** $t = T$**) do**
        $t \leftarrow t + 1$
        Observe current user u_t
        if $\ell \leq M_s$ **then** ▷ *Forced exploration*
            $a_t \leftarrow b_{p_{u_t}}$ ▷ *Round-robin over* $\mathcal{B}$
            $y_t \leftarrow$ UPDATE-EXPLORATION(t, u_t, a_t)
        else ▷ *Adaptive selection*
            if $\ell = M_s + 1$ **then** ▷ *Update graph once*
                $(\mathcal{G}_s, \{\mathcal{N}_s(i)\}_{i=1}^N) \leftarrow$ HM-EDGETEST$(\mathcal{G}_{s-1}, \{\boldsymbol{V}_{i,\mathrm{e},t-1}, \boldsymbol{b}_{i,\mathrm{e},t-1}, n_{i,\mathrm{e},t-1}\}_{i=1}^N, \lambda, \delta)$
            end if
            $\boldsymbol{V}_{\mathcal{N}_s(u_t),t-1} \leftarrow \lambda \boldsymbol{I}_d + \sum_{j \in \mathcal{N}_s(u_t)} (\boldsymbol{V}_{j,t-1} - \lambda \boldsymbol{I}_d)$
            $\boldsymbol{b}_{\mathcal{N}_s(u_t),t-1} \leftarrow \sum_{j \in \mathcal{N}_s(u_t)} \boldsymbol{b}_{j,t-1}$
            $\widehat{\boldsymbol{\theta}}_{\mathcal{N}_s(u_t),t-1} \leftarrow \boldsymbol{V}_{\mathcal{N}_s(u_t),t-1}^{-1} \boldsymbol{b}_{\mathcal{N}_s(u_t),t-1}$
            Choose $a_t \in \operatorname*{argmax}_{k \in [K]} \left\{ \boldsymbol{x}_k^\top \widehat{\boldsymbol{\theta}}_{\mathcal{N}_s(u_t),t-1} + \beta \|\boldsymbol{x}_k\|_{\boldsymbol{V}_{\mathcal{N}_s(u_t),t-1}^{-1}} \right\}$
            Play arm a_t and observe y_t
        end if
        $\boldsymbol{V}_{u_t,t} \leftarrow \boldsymbol{V}_{u_t,t-1} + \boldsymbol{x}_{a_t} \boldsymbol{x}_{a_t}^\top, \quad \boldsymbol{b}_{u_t,t} \leftarrow \boldsymbol{b}_{u_t,t-1} + y_t \boldsymbol{x}_{a_t}$
    end for
end for

4 ALGORITHM

In this section, we propose TRIBE in Algorithm 1. To encode the estimated cluster structure, TRIBE maintains a graph where each user is represented as a node, and each edge represents a relation between users. The graph is initialized as complete, so that every pair of nodes is connected, and is refined only by deleting edges. TRIBE maintains two sets of per-user statistics: exploration statistics $(\boldsymbol{V}_{i,\mathrm{e},t}, \boldsymbol{b}_{i,\mathrm{e},t}, n_{i,\mathrm{e},t})$ for cluster recovery and main statistics $(\boldsymbol{V}_{i,t}, \boldsymbol{b}_{i,t})$ for pooled UCB selection. Note that any statistic that is not updated at round t retains its value from round $t-1$.

Phase structure. As in PhaseUniCLUB (Li et al., 2025), TRIBE divides the horizon into phases $s = 0, 1, 2, \ldots$ of geometrically increasing length. Each phase s consists of a forced-exploration subphase of $M_s := C_0 2^s$ rounds followed by an adaptive-selection subphase of $L_s := \lceil C_0 2^{s/(1-p)} \rceil$ rounds, where C_0 denotes the phase-scale constant and $p \in (1/2, 1)$ controls how much faster the adaptive-selection subphases grow than the forced-exploration subphases. Let $h_s := \sum_{q=0}^{s-1} (M_q + L_q)$ denote the number of rounds before phase s, and let $t_s := h_s + M_s$ denote the last forced-exploration round of phase s. The graph is updated once per phase, at round $t_s + 1$, using the exploration statistics accumulated through round t_s (Line 19 of Algorithm 1), and every adaptive-selection round of phase s pools the main statistics over the resulting neighborhood $\mathcal{N}_s(u_t)$.

Round-robin block exploration. At initialization, TRIBE fixes an arbitrary ordering $(\boldsymbol{x}_{b_1}, \ldots, \boldsymbol{x}_{b_B})$ of a basis $\mathcal{B} \subseteq \mathcal{A}$ of W. In every forced-exploration round, the served user u_t is assigned the next basis arm in this order, tracked by a user-specific cyclic pointer $p_{u_t} \in [B]$ that persists across phases (Line 16). Each observation enters the main statistics immediately, whereas the exploration statistics of user i receive the observations only after the buffer $\mathcal{D}_i$ collects a complete block of B

Algorithm 2 `HM-EdgeTest` (Harmonic-Mean Edge Separation Test)

function HM-EDGETEST($\mathcal{G}_{s-1}, \{\boldsymbol{V}_{i,\mathrm{e},t-1}, \boldsymbol{b}_{i,\mathrm{e},t-1}, n_{i,\mathrm{e},t-1}\}_{i=1}^{N}, \lambda, \delta$)
  $\mathcal{G}_s \leftarrow \mathcal{G}_{s-1}$
  for each $i \in [N]$ **do**
    $\widehat{\boldsymbol{\theta}}_{i,\mathrm{e},t-1} \leftarrow \boldsymbol{V}_{i,\mathrm{e},t-1}^{-1} \boldsymbol{b}_{i,\mathrm{e},t-1}$
    $\beta_{i,\mathrm{e},t-1} \leftarrow \sqrt{2\log(N/\delta) + d\log\left(1 + \frac{n_{i,\mathrm{e},t-1}}{\lambda d}\right)} + \sqrt{\lambda}$
  end for
  for each edge $(i,j) \in E(\mathcal{G}_{s-1})$ **do**
    if $\|\widehat{\boldsymbol{\theta}}_{i,\mathrm{e},t-1} - \widehat{\boldsymbol{\theta}}_{j,\mathrm{e},t-1}\|_{\left(\boldsymbol{V}_{i,\mathrm{e},t-1}^{-1} + \boldsymbol{V}_{j,\mathrm{e},t-1}^{-1}\right)^{-1}} > \sqrt{\beta_{i,\mathrm{e},t-1}^2 + \beta_{j,\mathrm{e},t-1}^2}$ **then**
      Delete edge (i,j) from $\mathcal{G}_s$
    end if
  end for
  for each $i \in [N]$ **do**
    $\mathcal{N}_s(i) \leftarrow \{i\} \cup \{j : (i,j) \in E(\mathcal{G}_s)\}$
  end for
  return $\mathcal{G}_s$ and $\{\mathcal{N}_s(i)\}_{i=1}^{N}$
end function

Algorithm 3 `Update-Exploration`

function UPDATE-EXPLORATION(t, u_t, a_t)
  Play arm a_t and observe y_t
  Append $(\boldsymbol{x}_{a_t}, y_t)$ to $\mathcal{D}_{u_t}$
  $p_{u_t} \leftarrow (p_{u_t} \bmod B) + 1$
  if $|\mathcal{D}_{u_t}| = B$ **then**
    $\boldsymbol{V}_{u_t,\mathrm{e},t} \leftarrow \boldsymbol{V}_{u_t,\mathrm{e},t-1} + B\boldsymbol{\Sigma}_{\mathcal{B}}$
    $\boldsymbol{b}_{u_t,\mathrm{e},t} \leftarrow \boldsymbol{b}_{u_t,\mathrm{e},t-1} + \sum_{(\boldsymbol{x},y)\in\mathcal{D}_{u_t}} y\boldsymbol{x}$
    $n_{u_t,\mathrm{e},t} \leftarrow n_{u_t,\mathrm{e},t-1} + B$
    $\mathcal{D}_{u_t} \leftarrow \varnothing$
  end if
  return y_t
end function

observations. Since each completed block contains every basis arm exactly once, the same matrix $\sum_{h=1}^{B} \boldsymbol{x}_{b_h}\boldsymbol{x}_{b_h}^\top = B\boldsymbol{\Sigma}_{\mathcal{B}}$ is added to the exploration design matrix of the served user, $\boldsymbol{V}_{i,\mathrm{e},t}$. Let $E_i(t)$ denote the number of forced-exploration observations collected from user i through round t. Since at most $B-1$ of these observations remain in the incomplete buffer, the committed count equals $n_{i,\mathrm{e},t} = B\lfloor E_i(t)/B \rfloor$, and, for all $i \in [N]$, $\boldsymbol{V}_{i,\mathrm{e},t}$ takes the following common form.

Lemma 1 (Common structure of exploration statistics)**.** *For every user $i \in [N]$ and every round t,*

$$\boldsymbol{V}_{i,\mathrm{e},t} = \lambda \boldsymbol{I}_d + n_{i,\mathrm{e},t}\boldsymbol{\Sigma}_{\mathcal{B}}.$$

The proof of Lemma 1 is deferred to Appendix C.2.

Harmonic-mean separation. At each update, Algorithm 2 decides whether to delete each remaining edge (i,j). The test compares the exploration estimates $\widehat{\boldsymbol{\theta}}_{i,\mathrm{e},t_s}$ and $\widehat{\boldsymbol{\theta}}_{j,\mathrm{e},t_s}$ in the norm induced by the harmonic-mean matrix defined as $\boldsymbol{H}_{ij,t_s} := (\boldsymbol{V}_{i,\mathrm{e},t_s}^{-1} + \boldsymbol{V}_{j,\mathrm{e},t_s}^{-1})^{-1}$, which equals one half of the matrix harmonic mean of the two exploration design matrices. Our proposed statistical test based on $\boldsymbol{H}_{ij,t_s}$ gives a closed-form sufficient condition for the confidence ellipsoids to be disjoint, as presented in Lemma 2. In particular, by Lemma 4, the squared test statistic equals the smallest value of the combined squared distances $\|\widehat{\boldsymbol{\theta}}_{i,\mathrm{e},t_s} - \boldsymbol{z}\|_{\boldsymbol{V}_{i,\mathrm{e},t_s}}^2 + \|\widehat{\boldsymbol{\theta}}_{j,\mathrm{e},t_s} - \boldsymbol{z}\|_{\boldsymbol{V}_{j,\mathrm{e},t_s}}^2$ over all points $\boldsymbol{z} \in \mathbb{R}^d$, so no point can lie in both ellipsoids once the statistic exceeds the combined radius.

Lemma 2 (Separation test with harmonic mean)**.** *Suppose that the graph update in phase s is executed, and define $\mathcal{E}_{i,\mathrm{e},t_s} := \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{i,\mathrm{e},t_s}\|_{\boldsymbol{V}_{i,\mathrm{e},t_s}}^2 \le \beta_{i,\mathrm{e},t_s}^2\}$ for each $i \in [N]$. Then, for*

any $i, j \in [N]$, $\mathcal{E}_{i,\mathsf{e},t_s} \cap \mathcal{E}_{j,\mathsf{e},t_s} = \varnothing$ *if*

$$\left\| \widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s} - \widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s} \right\|_{\boldsymbol{H}_{ij,t_s}} > \sqrt{\beta^2_{i,\mathsf{e},t_s} + \beta^2_{j,\mathsf{e},t_s}}. \tag{4}$$

When Eq. (4) holds, Algorithm 2 deletes the edge (i, j). If $i, j \in \mathcal{I}_c$, then, with high probability, both confidence ellipsoids contain the shared observable parameter $\boldsymbol{\vartheta}^\star_c$, so the ellipsoids intersect, Eq. (4) fails, and the edge (i, j) survives. The following corollary formalizes this guarantee.

Corollary 1 (No false splits)**.** *With probability at least* $1 - \delta$ *for any fixed* $\delta \in (0, 1)$*, the separation test in Lemma 2 fails for all distinct users* $i, j \in [N]$ *with* $m(i) = m(j)$*. Hence every edge between users in the same observable cluster is preserved throughout the learning process.*

The detailed proofs of Lemma 2 and Corollary 1 are deferred to Appendix C.3.

Calibration and phase precision. `TRIBE` uses the phase-scale constant C_0 defined as

$$C_0 \coloneqq \lceil 16N \log(N/\delta) + 4Nn_0^\star \rceil, \quad n_0^\star \coloneqq B \left\lceil (2 + 2\sqrt{2})^2 \beta^2 / B \right\rceil,$$

where $\beta \coloneqq \sqrt{2\log(N/\delta) + d\log(1 + T/(\lambda d))} + \sqrt{\lambda}$ denotes the confidence coefficient for UCB selection. Here, $n_0^\star$ sets the baseline amount of exploration needed for a surviving edge to be informative. Consider any users $i, j \in [N]$ that have both committed at least $n_0^\star$ exploration samples. If the edge (i, j) survives the test in any phase s, the $\boldsymbol{\Sigma}_{\mathcal{B}}$-distance between their observable parameters is at most 1. Since only one user is served per round, $C_0 = \widetilde{\mathcal{O}}(Nd)$ makes the subphases long enough for all users to reach the baseline sample target $n_0^\star$ with high probability (see Lemma 8).

5 Main Results

5.1 Regret Analysis

Let $\rho_s \coloneqq 2^{-s/2}$ denote the precision level targeted at phase s for distinguishing observable parameters, and let $J_T \coloneqq 1 + \lceil \log_2(T + 1) \rceil$ denote an upper bound on the number of graph updates executed by time T. We define the first phase in which the conditions for exact recovery hold as

$$s_\pi \coloneqq \min \left\{ s \in \mathbb{Z}_{\geq 0} : N\pi_{\min} 2^s \geq 1 \text{ and } \rho_s < \Delta_{\mathcal{B}} \sqrt{N\pi_{\min}} \right\}.$$

The first condition, $N\pi_{\min} 2^s \geq 1$, ensures that every user accumulates sufficiently many committed exploration samples with high probability. Under the uniform user-arrival model, $N\pi_{\min} = 1$, so the first condition holds in every phase; the condition therefore accounts only for the slower sample accumulation of infrequently arriving users. The second condition requires the phase precision to go below the minimum observable cluster gap, ensuring that no cross-cluster edge can survive.

Theorem 1 (Exploration-driven observable recovery)**.** *Suppose Assumptions 1 and 2 hold and each user arrives following Eq.* (1)*. With probability at least* $1 - (J_T + 1)\delta$*, the following statements hold simultaneously at all graph updates executed by time* T*:*

(i) ***(Phasewise precision)*** *For every phase* s *satisfying* $N\pi_{\min} 2^s \geq 1$ *and every* $(i, j) \in E(\mathcal{G}_s)$*,*

$$\left\| \boldsymbol{\vartheta}^\star_{m(i)} - \boldsymbol{\vartheta}^\star_{m(j)} \right\|_{\boldsymbol{\Sigma}_{\mathcal{B}}} \leq \frac{\rho_s}{\sqrt{N\pi_{\min}}}.$$

(ii) ***(Exact recovery)*** *At every executed update in a phase* $s \geq s_\pi$*, the self-inclusive direct neighborhoods coincide with the observable clusters:*

$$\mathcal{N}_s(i) = \mathcal{I}_{m(i)} \qquad \textit{for every } i \in [N].$$

Theorem 1 guarantees, with high probability, in every phase where $N\pi_{\min} 2^s \geq 1$, that each surviving edge connects users whose observable parameters are within $\rho_s / \sqrt{N\pi_{\min}}$. Because the target precision decreases geometrically and users in different observable clusters are separated by at least $\Delta_{\mathcal{B}}$, all cross-cluster edges are eliminated from phase s_π onward. As all within-cluster edges are preserved, the resulting self-inclusive direct neighborhoods thereafter coincide with the observable clusters. The proof is presented in Appendix C.4.

We now establish the regret bound. Guided by Theorem 1, we decompose the cumulative regret as

$$\mathrm{Reg}(T) = \mathrm{Reg}_{\mathrm{post}}(T) + \mathrm{Reg}_{\mathrm{explore}}(T) + \mathrm{Reg}_{\mathrm{pre}}(T).$$

The post-recovery term $\mathrm{Reg}_{\mathrm{post}}(T)$ accumulates the regret of adaptive-selection rounds in phases $s \geq s_\pi$, the exploration term $\mathrm{Reg}_{\mathrm{explore}}(T)$ accumulates the regret of the forced-exploration rounds, and the identification term $\mathrm{Reg}_{\mathrm{pre}}(T)$ accumulates the regret of adaptive-selection rounds in phases $s < s_\pi$, before exact recovery is guaranteed.

Theorem 2 (Regret of TRIBE)**.** *Under the conditions of Theorem 1, with probability at least* $1 - (J_T + 2)\delta$*,*

$$\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(d\sqrt{CT} + (Nd)^p T^{1-p} + Nd(N\pi_{\min})^{-\frac{1}{1-p}} \Delta_{\mathcal{B}}^{-\frac{2}{1-p}} \right),$$

where the last term is omitted when $C = 1$*.*

$\mathrm{Reg}_{\mathrm{post}}(T) = \widetilde{\mathcal{O}}(d\sqrt{CT})$ matches the regret bound obtained by running OFUL (Abbasi-yadkori et al., 2011) separately within each observable cluster when the partition is known. For any $p \in (1/2, 1)$, $\mathrm{Reg}_{\mathrm{explore}}(T) = \widetilde{\mathcal{O}}((Nd)^p T^{1-p})$ grows strictly slower than $\sqrt{T}$, thus $\mathrm{Reg}_{\mathrm{post}}(T)$ asymptotically dominates. Note that the bound on $\mathrm{Reg}_{\mathrm{pre}}(T)$ depends on arrivals only through $N\pi_{\min} \in (0, 1]$, the ratio of the smallest arrival probability to $1/N$, because recovery requires even the least frequently arriving user to accumulate enough exploration samples; under uniform arrivals, $N\pi_{\min} = 1$ and the bound reduces to $\widetilde{\mathcal{O}}(Nd\Delta_{\mathcal{B}}^{-2/(1-p)})$. To the best of our knowledge, Theorem 2 provides the first regret bound of order $\widetilde{\mathcal{O}}(d\sqrt{CT})$ without stochastic feature diversity or an action set spanning $\mathbb{R}^d$, and the proof is provided in Appendix C.5.

5.2 Computational Analysis

Two design choices keep the computational cost of TRIBE low. First, since every committed exploration block contains each basis arm exactly once, all users' exploration design matrices share the same eigenbasis. Once the eigenbasis is computed, each pairwise separation test reduces to a sum over $B = \dim(W) \leq d$ coordinates, and the tests are run only at the $\mathcal{O}(\log T)$ graph updates.

Second, because the arm set is fixed, the pooled statistics $(\boldsymbol{V}_{\mathcal{N}_s(u_t),t}, \boldsymbol{b}_{\mathcal{N}_s(u_t),t})$ of the served user u_t are determined by the number of times each arm has been pulled by the users in $\mathcal{N}_s(u_t)$ up to round t and by the corresponding reward sums. We store the counts and reward sums separately for each arm, so each new observation updates $(\boldsymbol{V}_{\mathcal{N}_s(i),t}, \boldsymbol{b}_{\mathcal{N}_s(i),t})$ of every user $i \in \mathcal{N}_s(u_t)$ by increasing the count of arm a_t by one and adding the observed reward y_t to the corresponding reward sum, rather than adding a matrix and a vector. Each edge deletion likewise removes the past observations of the deleted neighbor by subtraction. As a result, excluding one-time preprocessing, the total arithmetic cost over T rounds is at most

$$\mathcal{O}\left(T(N + KB^2) + N^2K + N^2B\log(T+1)\right), \tag{5}$$

whereas BASS (Cherkaoui et al., 2025) costs $\mathcal{O}(T(Kd^2 + Nd^3 + N\tau d))$, since BASS runs an $\mathcal{O}(d^3 + \tau d)$ test with τ minimization steps on up to $N - 1$ edges in every round. The details and the derivation of Eq. (5) are provided in Appendix E.

6 Numerical Experiments

6.1 Experimental Setup

We evaluate a special case of our problem setting where $W = \mathbb{R}^d$, thus $C = C_{\mathrm{lat}}$, and user arrivals are uniform IID, to align with the baselines. Following the angle-parameterized construction of Cherkaoui et al. (2025), we construct $C = 3$ clusters with preference vectors $\boldsymbol{\theta}_1^\star = \boldsymbol{e}_1$, $\boldsymbol{\theta}_2^\star = \cos\phi\boldsymbol{e}_1 + \sin\phi\boldsymbol{e}_2$, and $\boldsymbol{\theta}_3^\star = \boldsymbol{e}_2$, where $\phi \in \{\pi/12, \pi/6, \pi/4\}$. Since all three vectors have unit norm, these angles map directly to Euclidean distances via $2\sin(\phi/2)$. In other words, ϕ alone sets the problem difficulty: the smaller ϕ, the harder it becomes to tell the pair $(\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2^\star)$ apart. In each run, we draw $K = 15$ arms IID from $\mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_d)$ with $d = 10$, normalize them to unit norm, and retain the resulting arm set throughout the horizon. We use $N = 100$ users with randomly permuted cluster memberships of sizes $(34, 34, 32)$, horizon $T = 8{,}000$, and IID reward noise $\epsilon_t \sim \mathcal{N}(0, 1)$.

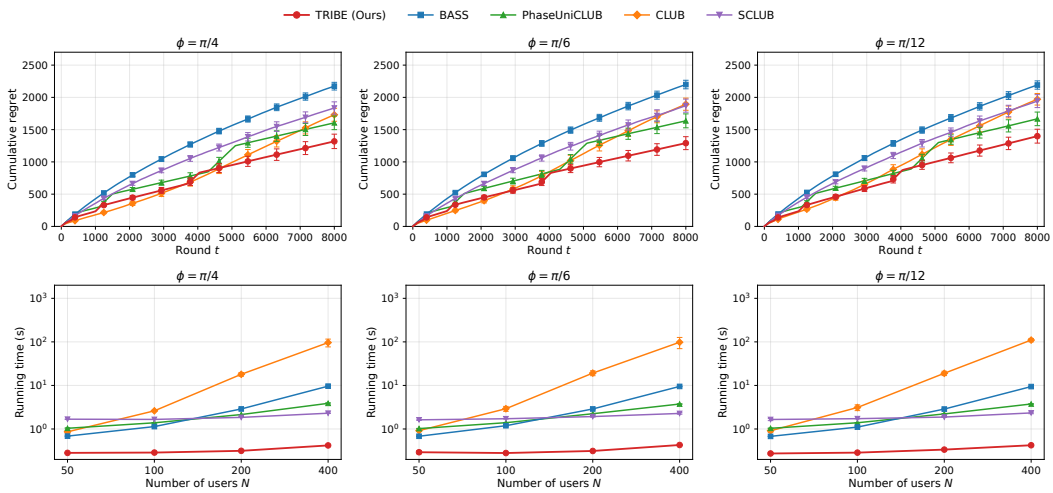


Figure 2: Cumulative regret (top) and runtime scalability (bottom) for $\phi = \pi/4$, $\pi/6$, and $\pi/12$ (left to right). Regret is measured over $T = 8{,}000$ rounds with $N = 100$ users, and runtime is measured at the same horizon with $N \in \{50, 100, 200, 400\}$; both are averaged over 5 seeds. Error bars indicate one standard error for regret and one standard deviation for runtime.

Baseline algorithms include CLUB (Gentile et al., 2014), SCLUB (Li et al., 2019), PhaseUniCLUB (Li et al., 2025), and BASS (Cherkaoui et al., 2025), where we adapt BASS to the one-user-per-round protocol. For TRIBE, the first $B = 10$ arms are chosen as the spanning basis in every run. We set $\lambda = 1$, $\delta = 0.05$, $p = 0.51$, and $C_0 = 10$. We report mean cumulative regret over 5 independent runs, with error bars representing one standard error of the mean.

For runtime scalability, we vary $N \in \{50, 100, 200, 400\}$ while keeping d, K, T, and the noise level fixed, and use the implementation described in Section 5.2. We report the mean wall-clock time of the interaction loop over 5 runs, excluding instance generation and initialization, with error bars representing one standard deviation. Experiments on real-world datasets are provided in Appendix B.

6.2 Experimental Results

Figure 2 reports cumulative regret (top) and runtime (bottom) for $\phi = \pi/4$, $\pi/6$, and $\pi/12$. Across all three regimes, TRIBE achieves the lowest cumulative regret over most of the horizon and at the terminal round. The regret curves of TRIBE and CLUB are initially similar, but their separation becomes clearer in later rounds. As ϕ decreases and the closest clusters become harder to distinguish, TRIBE continues to achieve lower regret than all baselines, including in the hardest regime $\phi = \pi/12$, indicating robustness of our algorithm to cluster separation.

The bottom panels show that the runtime of TRIBE increases only mildly over the tested population sizes, growing from $N = 50$ to $N = 400$ by less than a factor of two. TRIBE is the fastest algorithm at every population size, operating an order of magnitude faster than BASS and more than two orders of magnitude faster than CLUB. In contrast, the runtime of CLUB increases substantially between $N = 200$ and $N = 400$. The similarity of runtime curves across the three angular regimes suggests that computational cost is determined primarily by population size rather than cluster separation.

7 Conclusion

We studied online clustering of bandits through observable clusters defined by mean-reward equivalence rather than parameter identity. Our algorithm, TRIBE, combines spanning-basis exploration, harmonic-mean separation, and neighborhood-pooled UCB, dispensing with the restrictive assumptions required for parameter identification in prior studies. TRIBE achieves a regret bound with a leading term of $\widetilde{\mathcal{O}}(d\sqrt{CT})$, matching the rate attainable when the cluster structure is known a priori. Moreover, the block exploration strategy gives the design matrices a common eigenbasis, enabling $\mathcal{O}(d)$ pairwise separation tests. Empirical results support our theoretical analyses in terms of both regret performance and computational efficiency.

AI USE STATEMENT

In this work, we used generative AI tools to edit the manuscript for readability, format references, and debug code for the numerical experiments. We did not use them for any research task for which ICLR requires disclosure, including developing theoretical models, formulating or proving mathematical claims, proposing hypotheses, designing methodology or experiments, implementing our proposed methods, generating or cleaning data, translation, and interpreting results. In particular, all parts constituting this paper – problem formulations, theoretical results, algorithm designs, and empirical conclusions – are the authors' original work. We have reviewed all AI-assisted work: every AI-assisted code change was read and independently re-run by the authors, and all AI-assisted edits were checked against the original text to ensure no claim or citation was altered. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 127–135. PMLR, 2013.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

Yikun Ban and Jingrui He. Local clustering in contextual multi-armed bandits. In *Proceedings of the Web Conference 2021*, WWW '21, pp. 2335–2346. Association for Computing Machinery, 2021.

Nicolò Cesa-Bianchi, Claudio Gentile, and Giovanni Zappella. A gang of bandits. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, pp. 737–745. Curran Associates Inc., 2013.

Hamza Cherkaoui, Merwan Barlier, and Igor Colin. Adaptive sample sharing for multi agent linear bandits. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10216–10248. PMLR, 2025.

Anh Do, Thanh Nguyen-Tang, and Raman Arora. Multi-agent learning with heterogeneous linear contextual bandits. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2023.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 757–765. PMLR, 2014.

Claudio Gentile, Shuai Li, Purushottam Kar, Alexandros Karatzoglou, Giovanni Zappella, and Evans Etrue. On context-dependent clustering of bandits. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1253–1262. PMLR, 2017.

Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001. doi: 10.1023/A:1011419012209.

F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19:1–19:19, 2015. doi: 10.1145/2827872.

Jiafan He, Tianhao Wang, Yifei Min, and Quanquan Gu. A simple and provably efficient algorithm for asynchronous federated contextual linear bandits. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4762–4775. Curran Associates, Inc., 2022.

Tor Lattimore and Csaba Szepesvari. The End of Optimism? An Asymptotic Analysis of Finite-Armed Linear Bandits. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 728–737. PMLR, 2017.

Chung-Wei Lee, Qinghua Liu, Yasin Abbasi Yadkori, Chi Jin, Tor Lattimore, and Csaba Szepesvari. Context-lumpable stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77915–77941. Curran Associates, Inc., 2023.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 661–670. Association for Computing Machinery, 2010.

Shuai Li and Shengyu Zhang. Online clustering of contextual cascading bandits. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

Shuai Li, Alexandros Karatzoglou, and Claudio Gentile. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pp. 539–548. Association for Computing Machinery, 2016.

Shuai Li, Wei Chen, Shuai Li, and Kwong-Sak Leung. Improved algorithm on online clustering of bandits. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 2923–2929. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

Zhuohua Li, Maoli Liu, Xiangxiang Dai, and John C.S. Lui. Demystifying online clustering of bandits: Enhanced exploration under stochastic and smoothed adversarial contexts. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xutong Liu, Haoru Zhao, Tong Yu, Shuai Li, and John C.S. Lui. Federated online clustering of bandits. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1221–1231. PMLR, 01–05 Aug 2022.

Sabina Tomkins, Peng Liao, Predrag Klasnja, and Susan Murphy. IntelligentPooling: practical Thompson sampling for mhealth. *Machine Learning*, 110(9):2685–2727, 2021.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Zhiyong Wang, Jize Xie, Xutong Liu, Shuai Li, and John C.S. Lui. Online clustering of bandits with misspecified user models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 3785–3818. Curran Associates, Inc., 2023a.

Zhiyong Wang, Jize Xie, Tong Yu, Shuai Li, and John C.S. Lui. Online corrupted user detection and regret minimization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 33262–33287. Curran Associates, Inc., 2023b.

Zhiyong Wang, Jiahang Sun, Mingze Kong, Jize Xie, Qinghua Hu, John C.S. Lui, and Zhongxiang Dai. Online clustering of dueling bandits. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 64403–64431. PMLR, 13–19 Jul 2025.

Hantao Yang, Xutong Liu, Zhiyong Wang, Hong Xie, John C. S. Lui, Defu Lian, and Enhong Chen. Federated contextual cascading bandits with asynchronous communication and heterogeneous users. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024.

Junwen Yang and Vincent Tan. Minimax optimal fixed-budget best arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, volume 35, pp. 12253–12266. Curran Associates, Inc., 2022.

Kaige Yang, Laura Toni, and Xiaowen Dong. Laplacian-regularized graph bandits: Algorithms and theoretical analysis. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3133–3143. PMLR, 2020a.

Liu Yang, Bo Liu, Leyu Lin, Feng Xia, Kai Chen, and Qiang Yang. Exploring clustering of bandits for online recommendation system. In *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, pp. 120–129, New York, NY, USA, 2020b. Association for Computing Machinery.

Yelp Inc. Yelp open dataset. https://www.yelp.com/dataset, 2024. Accessed: 2026-09-23.

A Extended Related Work

Information sharing under user heterogeneity. Information sharing across multiple users has been studied through Laplacian regularization (Cesa-Bianchi et al., 2013; Yang et al., 2020a), heterogeneity-aware multi-agent collaboration (Do et al., 2023), and federated bandit formulations for communication-efficient information sharing across decentralized agents (He et al., 2022; Liu et al., 2022; Yang et al., 2024). The online clustering-of-bandits literature, introduced by Gentile et al. (2014), models user heterogeneity through an unknown similarity structure over users and jointly learns this structure while selecting arms using information shared across related users. Extensions include user-item clustering (Li et al., 2016), arm-dependent user grouping (Gentile et al., 2017), cascading feedback (Li & Zhang, 2018), set-based cluster representation (Li et al., 2019), mixed-membership user models (Yang et al., 2020b), misspecified or corrupted user models (Wang et al., 2023a;b), and pairwise preference feedback (Wang et al., 2025).

Grouping criteria and statistical guarantees. Many clustering-of-bandits methods use parameter identity, grouping users whose preference vectors have the same values in all coordinates (Gentile et al., 2014; Li & Zhang, 2018; Li et al., 2019; Liu et al., 2022). Other methods let the user grouping depend on the arm (Li et al., 2016; Gentile et al., 2017): Li et al. (2016) also group arms according to whether the arms induce the same grouping of users, while Gentile et al. (2017) estimate user neighborhoods separately for each arm. Ban & He (2021) replace exact equality by local similarity, allowing bounded parameter differences within overlapping local clusters. Context-lumpable bandits (Lee et al., 2023) group users whose expected rewards are the same across all actions, in a finite user-action reward table without linear action features; for S users, K actions, and r groups, their online algorithm requires r as input and attains a gap-free expected-regret bound of $\widetilde{\mathcal{O}}(\sqrt{r^3(S+K)T})$, which corresponds to $\widetilde{\mathcal{O}}(\sqrt{C^3(N+K)T})$ in our setting with $S=N$ users, K arms, and $r=C$ groups. Our observable clusters likewise group users whose expected rewards are the same for every available arm, but exploit the linear feature structure of the arms: two users belong to the same observable cluster precisely when their preference-vector difference is orthogonal to every available arm. Unlike approximate parameter similarity, membership in an observable cluster preserves the expected-reward model exactly even when the full preference vectors differ. Compared with context-lumpable bandits, `TRIBE` requires no knowledge of the number C of observable clusters, and the leading term $\widetilde{\mathcal{O}}(d\sqrt{CT})$ of the regret bound depends, up to logarithmic factors, on neither the number of users nor the number of actions.

Recovering user groups requires sufficient information to distinguish users according to the chosen grouping criterion. Earlier analyses obtain such information through stochastic feature diversity and additional variance-proxy restrictions (Gentile et al., 2014; Li & Zhang, 2018; Li et al., 2019; Liu et al., 2022), which control the growth of each user's design matrix under adaptive UCB selection so that confidence regions contract in every direction of $\mathbb{R}^d$ without an explicit forced-exploration schedule. To remove the variance-proxy restrictions, Li et al. (2025) introduce forced uniform exploration while retaining IID feature generation and a positive-definite population second-moment matrix; the gap-agnostic algorithm `PhaseUniCLUB` combines the exploration with a phase schedule and direct-neighbor pooling. Moreover, Li et al. (2025) also study smoothed-adversarial variants, while random feature perturbations using Gaussian noise provide diversity again. Cherkaoui et al. (2025) dispense with stochastic feature diversity in their algorithm, `BASS`, but assume an arm set spanning the entire feature space $\mathbb{R}^d$. `TRIBE` instead obtains the information needed for observable clusters through repeated exploration over a basis of the available arms, commits exploration observations to the separation tests only in complete blocks, and retains full-history statistics for UCB selection, so that the main setting requires neither stochastic feature diversity nor an arm set spanning the entire feature space $\mathbb{R}^d$. In Appendix D, we extend our analysis to time-varying features, which requires IID feature generation only; the positive-definite second-moment matrix assumption is not necessary.

`BASS` also serves every user in each iteration, and the clustered-model guarantee of `BASS` is stated per user with the sharing factor $1/\sqrt{\rho_{\min}N}$, where $\rho_{\min}$ denotes the minimum cluster proportion. In contrast, our guarantee bounds cumulative regret over T interactions with one user per round under unknown non-uniform IID user arrivals, with a pooled-learning term that matches the corresponding `OFUL` benchmark when the observable clusters are known and requires no balance among observable cluster sizes.

Comparison with context-dependent clustering. The arm-dependent user grouping of Gentile et al. (2017) also compares users through expected rewards, and TRIBE differs from the approach in three respects. First, the regret bound established by Gentile et al. (2017) includes a data-dependent hardness term, defined as the maximum number of rounds required for each user's design matrix, thus measuring the worst-case delay, to accumulate information in every direction. Their analysis controls the hardness term under stochastic feature diversity and an additional variance restriction, as in Gentile et al. (2014), thereby obtaining a regret bound expressed in terms of the expected number of clusters. In the absence of these assumptions, the hardness term may increase linearly with T, as the authors observe. In particular, when the available arm set does not span $\mathbb{R}^d$, the design matrices never accumulate information in the directions orthogonal to every available arm, no matter how extensively the arms are explored, so the hardness term grows linearly with T. Therefore, keeping the action set unchanged does not by itself resolve the difficulty of cluster recovery. The stochastic assumptions in prior analyses ensure that identifying information accumulates, whereas TRIBE must acquire the information needed for observable-cluster recovery through structured exploration. By combining spanning-basis exploration with harmonic-mean tests, TRIBE achieves a leading regret term of $\widetilde{\mathcal{O}}(d\sqrt{CT})$ without requiring stochastic feature diversity or an action set spanning the entire feature space.

Second, their algorithm, CAB, requires a separation parameter as input. This parameter establishes a lower bound on the expected reward gap between users in different neighborhoods of the same arm and determines when observations are shared. In contrast, TRIBE does not require any knowledge of the gap between observable clusters. Third, CAB recomputes the served user's neighborhood by comparing against all users for every available arm in every round. In contrast, TRIBE maintains a single user graph shared across all arms and updates the graph only at $\mathcal{O}(\log(T+1))$ phase transitions, avoiding neighborhood re-estimation between updates.

Pairwise separation and computational cost. The choice of pairwise separation criterion also affects computational cost. BASS tests whether two confidence ellipsoids intersect by an exact criterion (Cherkaoui et al., 2025). The synchronous protocol of BASS makes the design matrices of the compared users identical, but when one user is served per round, the design matrices differ across users, and the pairwise implementation reported for BASS then requires matrix decompositions and one-dimensional numerical minimization costing $\mathcal{O}(d^3 + \tau d)$ per pair, where τ is the number of optimization iterations; moreover, the test is applied in every round to each of the up to $N-1$ edges incident to the served user. TRIBE instead uses the harmonic-mean test, a closed-form sufficient condition for disjointness that still preserves every within-cluster edge (Corollary 1), and applies the test only at the $\mathcal{O}(\log T)$ phase transitions.

In the main setting, the exploration design of TRIBE further removes pairwise decompositions. TRIBE fixes an arbitrary ordering of a basis $\mathcal{B} \subseteq \mathcal{A}$ of W, with $B = |\mathcal{B}| = \dim(W) \leq d$, and commits each user's exploration observations only after completing a block containing every basis arm exactly once. Each completed block therefore contributes the same matrix $B\boldsymbol{\Sigma}_{\mathcal{B}}$ to the user's exploration design, where $\boldsymbol{\Sigma}_{\mathcal{B}} := B^{-1}\sum_{\boldsymbol{x}\in\mathcal{B}} \boldsymbol{x}\boldsymbol{x}^\top$. Consequently, at every graph update based on observations through t_s, each exploration design matrix $\boldsymbol{V}_{i,\mathtt{e},t_s}$ equals the common ridge term $\lambda \boldsymbol{I}_d$ plus $\boldsymbol{\Sigma}_{\mathcal{B}}$ scaled by the committed exploration count $n_{i,\mathtt{e},t_s}$ (Lemma 1). The exploration designs therefore share a common eigenbasis even when users have committed different numbers of observations, unlike the identical designs that BASS obtains only through synchronous serving. After one-time preprocessing, the harmonic-mean statistic can then be evaluated coordinate-wise in $\mathcal{O}(B) \leq \mathcal{O}(d)$ time per pair, without pair-specific eigendecomposition or iterative numerical minimization. The comparison concerns pairwise test evaluation; the total computational cost, including pooled action selection and historical corrections after edge deletions, is established in Section 5.2 and Appendix E.

B Real-world Experiments

B.1 Datasets

We evaluate on three user–item interaction datasets. MovieLens-100K (Harper & Konstan, 2015) contains 100,000 ratings on a $[5] = \{1, 2, 3, 4, 5\}$ integer scale from 943 users on 1,682 movies. Jester (Goldberg et al., 2001) contains continuous ratings in $[-10, 10]$ from 73,421 users on 100 jokes. The Yelp Open Dataset (Yelp Inc., 2024) contains business reviews with star ratings in $[5]$; we use

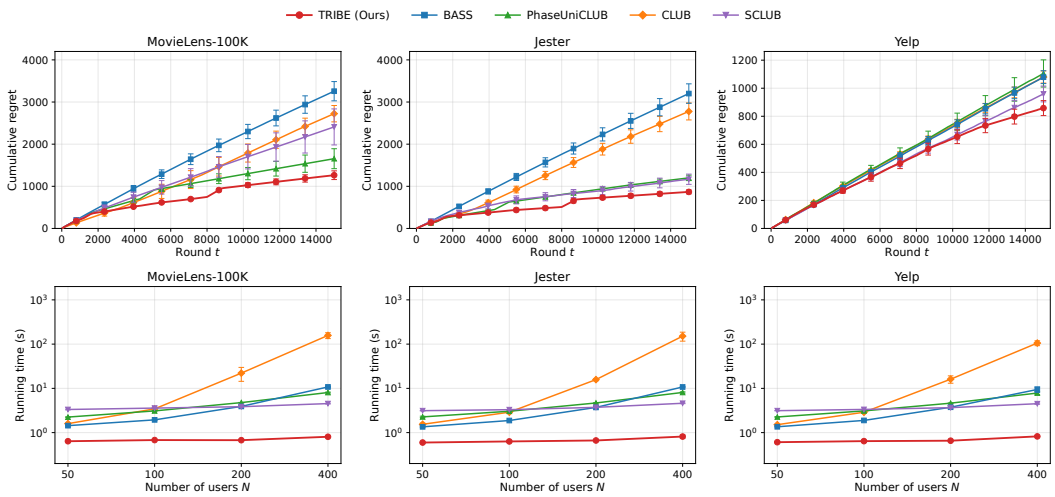


Figure 3: Cumulative regret (top) and runtime scalability (bottom) on MovieLens-100K, Jester, and Yelp (left to right). Regret is measured over $T = 15{,}000$ rounds, and runtime is measured with $N \in \{50, 100, 200, 400\}$ users; both are averaged over 5 seeds. Error bars indicate one standard error for regret and one standard deviation for runtime.

the New Orleans, LA, subset, comprising 635,521 reviews from 245,421 users on 6,215 businesses, since a fixed arm set shared by all users is well-posed only within one metropolitan area.

B.2 Experimental Setup

All three datasets use the same data-derived linear-bandit pipeline, following the real-data protocol of Li et al. (2025). We retain the 200 most active users and the 1,500 items with the most interactions; for Jester, whose catalog is smaller, we retain all jokes. Ratings are binarized at the midpoint of the corresponding scale: an observed entry equals one when the rating exceeds 3 for MovieLens and Yelp or exceeds 0 for Jester, and zero otherwise. For the resulting binary matrix $\boldsymbol{R}$, a singular value decomposition $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ provides user taste vectors $\widetilde{\boldsymbol{\theta}}_i = (\boldsymbol{U}\boldsymbol{S})_{i,1:10}$ and item features from the corresponding rows of $\boldsymbol{V}$. Items with zero mass in this $d = 10$ subspace are discarded, and each remaining item feature is normalized to unit ℓ_2-norm.

For each seed, we sample 150 users uniformly from the retained 200 users and partition their taste vectors into $C = 10$ groups using k-means with 50 iterations. If an empty group occurs, one randomly chosen user is assigned to it. Each cluster parameter is then set to the mean taste vector of its members, all centroids are jointly rescaled so that the largest centroid norm equals one, and every user receives its cluster centroid as the ground-truth parameter. The arm set contains $K = 50$ items sampled uniformly without replacement from the retained catalog and redrawn until it has rank d; this arm set remains fixed throughout the run. Hence $W = \mathbb{R}^d$, and the observable and latent partitions coincide, with $C = C_{\text{lat}} = 10$ and the unbalanced cluster sizes generated by k-means.

Each run has $N = 150$ users and horizon $T = 15{,}000$. Users arrive uniformly, so $\pi_i = 1/N$ for every $i \in [N]$. After an arm is selected, the reward follows Eq. (2) with independent noise $\epsilon_t \sim \mathcal{N}(0, 1)$. We compare `TRIBE` with the same four baselines as in the synthetic experiments (see Section 6.1). For `TRIBE`, we set $\lambda = 1$, $\delta = 0.05$, $p = 2/3$, and $C_0 = 100$. For the running-time comparison, we vary $N \in \{50, 100, 200, 400\}$. For each N, we retain the $\max\{N, 200\}$ most active users in place of 200 and sample N of them uniformly without replacement, keeping all other settings fixed. We report the mean wall-clock time of a full run over 5 seeds with one standard deviation.

B.3 Experimental Results

Figure 3 presents cumulative regret (top) and runtime scalability (bottom) for the MovieLens-100K, Jester, and Yelp datasets. Across all three datasets, `TRIBE` achieves the lowest level of cumulative regret. On MovieLens, `TRIBE` outperforms all baselines, with `PhaseUniCLUB` as the

closest competitor. On Jester, the advantage of TRIBE over CLUB and BASS becomes increasingly pronounced as the horizon extends. Although PhaseUniCLUB and SCLUB remain closer, they still exhibit higher regret than TRIBE. On Yelp, the methods demonstrate more similar performance; however, TRIBE maintains the lowest mean regret during the later rounds and at $T = 15{,}000$. Therefore, the outperformance of our algorithm, in terms of cumulative regret, observed in synthetic datasets persists across all three instances derived from real-world datasets.

The runtime panels indicate that TRIBE achieves the lowest mean runtime across all tested population sizes and datasets. The runtime of TRIBE increases only modestly as N increases from 50 to 400, whereas CLUB exhibits the steepest increase. The runtimes of BASS and PhaseUniCLUB also increase substantially with N, while SCLUB grows more gradually but remains slower than TRIBE. This behavior aligns with the scalar-propagation implementation and coordinate-wise harmonic-mean tests described in Section 5.2. The reported regret curves assess decision quality but do not directly measure recovery of the observable partition.

C Missing Proofs

C.1 Relation Between Latent and Observable Partitions

Before providing the proofs, we first clarify the relation between the latent and observable partitions. The partition we seek to recover groups users according to equality of their projected parameters rather than their parameter identities. The following lemma characterizes the resulting relation between the two partitions.

Lemma 3 (Relation between latent and observable partitions)**.** *Each observable cluster is a union of latent clusters, and hence $C \le C_{\mathrm{lat}}$, where equality holds if and only if distinct latent parameters have distinct projections onto W.*

Proof of Lemma 3. Let $\Theta := \{\boldsymbol{\theta}_i^\star : i \in [N]\}$ denote the set of distinct preference vectors, so that $C_{\mathrm{lat}} = |\Theta|$, and for each $\boldsymbol{\theta} \in \Theta$, let $\mathcal{L}_{\boldsymbol{\theta}} := \{i \in [N] : \boldsymbol{\theta}_i^\star = \boldsymbol{\theta}\}$ denote the corresponding latent cluster. Let $\Pi_W(\Theta) := \{\Pi_W \boldsymbol{\theta} : \boldsymbol{\theta} \in \Theta\}$ denote the image of Θ under Π_W. By Definition 1, the observable clusters correspond one-to-one to the distinct observable parameters, that is, $\Pi_W(\Theta) = \{\boldsymbol{\vartheta}_c^\star : c \in [C]\}$, so that $C = |\Pi_W(\Theta)|$ and $\mathcal{I}_c = \{i \in [N] : \Pi_W \boldsymbol{\theta}_i^\star = \boldsymbol{\vartheta}_c^\star\}$ for each $c \in [C]$.

We first show that, for every $c \in [C]$,

$$\mathcal{I}_c = \bigcup_{\boldsymbol{\theta} \in \Theta:\, \Pi_W \boldsymbol{\theta} = \boldsymbol{\vartheta}_c^\star} \mathcal{L}_{\boldsymbol{\theta}}. \tag{6}$$

If $i \in \mathcal{I}_c$, then $\boldsymbol{\theta}_i^\star \in \Theta$ satisfies $\Pi_W \boldsymbol{\theta}_i^\star = \boldsymbol{\vartheta}_c^\star$, and $i \in \mathcal{L}_{\boldsymbol{\theta}_i^\star}$, which is one of the sets in the union.

Conversely, if $i \in \mathcal{L}_{\boldsymbol{\theta}}$ for some $\boldsymbol{\theta} \in \Theta$ with $\Pi_W \boldsymbol{\theta} = \boldsymbol{\vartheta}_c^\star$, then $\boldsymbol{\theta}_i^\star = \boldsymbol{\theta}$, so $\Pi_W \boldsymbol{\theta}_i^\star = \boldsymbol{\vartheta}_c^\star$ and $i \in \mathcal{I}_c$. Since the latent clusters are pairwise disjoint, the union in Eq. (6) is disjoint. Thus each observable cluster is a disjoint union of latent clusters, and the observable partition is a coarsening of the latent partition.

Next, consider the restriction of Π_W to Θ, that is, the map $\boldsymbol{\theta} \mapsto \Pi_W \boldsymbol{\theta}$ from Θ to $\Pi_W(\Theta)$. This map is surjective by the definition of $\Pi_W(\Theta)$, so

$$C = |\Pi_W(\Theta)| \le |\Theta| = C_{\mathrm{lat}}.$$

Since Θ is finite, a surjection from Θ onto $\Pi_W(\Theta)$ is injective if and only if $|\Pi_W(\Theta)| = |\Theta|$. Hence $C = C_{\mathrm{lat}}$ if and only if $\Pi_W \boldsymbol{\theta} \neq \Pi_W \boldsymbol{\theta}'$ for all distinct $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, that is, if and only if distinct latent parameters have distinct projections onto W. By Eq. (6), this condition is equivalent to every observable cluster consisting of exactly one latent cluster. □

C.2 Proof of Lemma 1

Fix a user $i \in [N]$ and a round $t \in \{0, \dots, T\}$, and recall that $E_i(t)$ denotes the number of forced-exploration observations collected from user i through round t. The pointer p_i starts at 1 with an empty buffer $\mathcal{D}_i$, advances cyclically through the ordered basis $(\boldsymbol{x}_{b_1}, \dots, \boldsymbol{x}_{b_B})$ at every forced-exploration round serving user i, and, together with the buffer, persists across phase boundaries. The buffer is

emptied exactly when it holds B observations (Lines 5 and 9 of Algorithm 3), at which point the pointer has returned to 1. Consequently, every complete block consists of B consecutive exploration observations of user i that contain each basis arm exactly once, and the first $E_i(t)$ exploration observations consist of

$$q_i(t) \coloneqq \left\lfloor \frac{E_i(t)}{B} \right\rfloor$$

complete blocks and fewer than B observations in the current incomplete block. Only complete blocks are committed to the exploration statistics, so the observations remaining in the buffer do not yet contribute to $\boldsymbol{V}_{i,\mathsf{e},t}$.

By the definition of $\boldsymbol{\Sigma}_{\mathcal{B}}$, the contribution of each complete block to the exploration design matrix is

$$\sum_{\ell=1}^{B} \boldsymbol{x}_{b_\ell} \boldsymbol{x}_{b_\ell}^\top = B\boldsymbol{\Sigma}_{\mathcal{B}}.$$

Since $\boldsymbol{V}_{i,\mathsf{e},0} = \lambda \boldsymbol{I}_d$ and each of the $q_i(t)$ committed blocks adds the same increment $B\boldsymbol{\Sigma}_{\mathcal{B}}$, we obtain

$$\begin{aligned} \boldsymbol{V}_{i,\mathsf{e},t} &= \lambda \boldsymbol{I}_d + q_i(t) B \boldsymbol{\Sigma}_{\mathcal{B}} \\ &= \lambda \boldsymbol{I}_d + n_{i,\mathsf{e},t} \boldsymbol{\Sigma}_{\mathcal{B}}, \end{aligned}$$

where the last equality holds because each commit increases the committed count by B (Line 8 of Algorithm 3), so that $n_{i,\mathsf{e},t} = Bq_i(t) = B\lfloor E_i(t)/B \rfloor$. □

C.3 Proofs of Lemma 2 and Corollary 1

Lemma 4 (Harmonic-mean quadratic identity)**.** *For symmetric positive-definite matrices $\boldsymbol{C}, \boldsymbol{D}$ and vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, let $\boldsymbol{H} \coloneqq (\boldsymbol{C}^{-1} + \boldsymbol{D}^{-1})^{-1}$. Then*

$$\|\boldsymbol{u} - \boldsymbol{v}\|_{\boldsymbol{H}}^2 = \min_{\boldsymbol{z} \in \mathbb{R}^d} \left\{ \|\boldsymbol{u} - \boldsymbol{z}\|_{\boldsymbol{C}}^2 + \|\boldsymbol{v} - \boldsymbol{z}\|_{\boldsymbol{D}}^2 \right\}.$$

Proof of Lemma 4. Define

$$f(\boldsymbol{z}) \coloneqq \|\boldsymbol{u} - \boldsymbol{z}\|_{\boldsymbol{C}}^2 + \|\boldsymbol{v} - \boldsymbol{z}\|_{\boldsymbol{D}}^2.$$

Expanding the two quadratic terms gives

$$f(\boldsymbol{z}) = \boldsymbol{u}^\top \boldsymbol{C}\boldsymbol{u} + \boldsymbol{v}^\top \boldsymbol{D}\boldsymbol{v} - 2\boldsymbol{z}^\top (\boldsymbol{C}\boldsymbol{u} + \boldsymbol{D}\boldsymbol{v}) + \boldsymbol{z}^\top (\boldsymbol{C} + \boldsymbol{D})\boldsymbol{z}. \tag{7}$$

Because $\boldsymbol{C} + \boldsymbol{D} \succ \boldsymbol{0}_{d\times d}$, the Hessian $2(\boldsymbol{C} + \boldsymbol{D})$ is positive definite, so f is strictly convex and its unique minimizer is the stationary point. Setting the gradient ∇f equal to zero gives

$$(\boldsymbol{C} + \boldsymbol{D})\boldsymbol{z}_\star = \boldsymbol{C}\boldsymbol{u} + \boldsymbol{D}\boldsymbol{v},$$

and therefore

$$\boldsymbol{z}_\star = (\boldsymbol{C} + \boldsymbol{D})^{-1}(\boldsymbol{C}\boldsymbol{u} + \boldsymbol{D}\boldsymbol{v}).$$

Substituting $\boldsymbol{z}_\star$ into Eq. (7) yields

$$f(\boldsymbol{z}_\star) = \boldsymbol{u}^\top \boldsymbol{C}\boldsymbol{u} + \boldsymbol{v}^\top \boldsymbol{D}\boldsymbol{v} - (\boldsymbol{C}\boldsymbol{u} + \boldsymbol{D}\boldsymbol{v})^\top (\boldsymbol{C} + \boldsymbol{D})^{-1}(\boldsymbol{C}\boldsymbol{u} + \boldsymbol{D}\boldsymbol{v}). \tag{8}$$

To express Eq. (8) in terms of $\boldsymbol{u} - \boldsymbol{v}$, use $\boldsymbol{C}\boldsymbol{u} + \boldsymbol{D}\boldsymbol{v} = (\boldsymbol{C} + \boldsymbol{D})\boldsymbol{v} + \boldsymbol{C}(\boldsymbol{u} - \boldsymbol{v})$. Expanding the resulting quadratic form and canceling all terms involving only $\boldsymbol{v}$ gives

$$\begin{aligned} f(\boldsymbol{z}_\star) &= \boldsymbol{u}^\top \boldsymbol{C}\boldsymbol{u} + \boldsymbol{v}^\top \boldsymbol{D}\boldsymbol{v} - \left[(\boldsymbol{C} + \boldsymbol{D})\boldsymbol{v} + \boldsymbol{C}(\boldsymbol{u} - \boldsymbol{v})\right]^\top (\boldsymbol{C} + \boldsymbol{D})^{-1} \left[(\boldsymbol{C} + \boldsymbol{D})\boldsymbol{v} + \boldsymbol{C}(\boldsymbol{u} - \boldsymbol{v})\right] \\ &= \left[\boldsymbol{v} + (\boldsymbol{u} - \boldsymbol{v})\right]^\top \boldsymbol{C} \left[\boldsymbol{v} + (\boldsymbol{u} - \boldsymbol{v})\right] + \boldsymbol{v}^\top \boldsymbol{D}\boldsymbol{v} \\ &\quad - \left[(\boldsymbol{C} + \boldsymbol{D})\boldsymbol{v} + \boldsymbol{C}(\boldsymbol{u} - \boldsymbol{v})\right]^\top (\boldsymbol{C} + \boldsymbol{D})^{-1} \left[(\boldsymbol{C} + \boldsymbol{D})\boldsymbol{v} + \boldsymbol{C}(\boldsymbol{u} - \boldsymbol{v})\right] \\ &= \boldsymbol{v}^\top \boldsymbol{C}\boldsymbol{v} + 2\boldsymbol{v}^\top \boldsymbol{C}(\boldsymbol{u} - \boldsymbol{v}) + (\boldsymbol{u} - \boldsymbol{v})^\top \boldsymbol{C}(\boldsymbol{u} - \boldsymbol{v}) + \boldsymbol{v}^\top \boldsymbol{D}\boldsymbol{v} \\ &\quad - \left[\boldsymbol{v}^\top (\boldsymbol{C} + \boldsymbol{D})\boldsymbol{v} + \boldsymbol{v}^\top \boldsymbol{C}(\boldsymbol{u} - \boldsymbol{v}) + (\boldsymbol{u} - \boldsymbol{v})^\top \boldsymbol{C}\boldsymbol{v} + (\boldsymbol{u} - \boldsymbol{v})^\top \boldsymbol{C}(\boldsymbol{C} + \boldsymbol{D})^{-1}\boldsymbol{C}(\boldsymbol{u} - \boldsymbol{v})\right] \\ &= \boldsymbol{v}^\top (\boldsymbol{C} + \boldsymbol{D})\boldsymbol{v} + 2\boldsymbol{v}^\top \boldsymbol{C}(\boldsymbol{u} - \boldsymbol{v}) + (\boldsymbol{u} - \boldsymbol{v})^\top \boldsymbol{C}(\boldsymbol{u} - \boldsymbol{v}) \\ &\quad - \left[\boldsymbol{v}^\top (\boldsymbol{C} + \boldsymbol{D})\boldsymbol{v} + 2\boldsymbol{v}^\top \boldsymbol{C}(\boldsymbol{u} - \boldsymbol{v}) + (\boldsymbol{u} - \boldsymbol{v})^\top \boldsymbol{C}(\boldsymbol{C} + \boldsymbol{D})^{-1}\boldsymbol{C}(\boldsymbol{u} - \boldsymbol{v})\right] \\ &= (\boldsymbol{u} - \boldsymbol{v})^\top \left[\boldsymbol{C} - \boldsymbol{C}(\boldsymbol{C} + \boldsymbol{D})^{-1}\boldsymbol{C}\right](\boldsymbol{u} - \boldsymbol{v}) \\ &= (\boldsymbol{u} - \boldsymbol{v})^\top \boldsymbol{C}(\boldsymbol{C} + \boldsymbol{D})^{-1}\boldsymbol{D}(\boldsymbol{u} - \boldsymbol{v}), \end{aligned}$$

where the last equality follows from

$$\boldsymbol{C} - \boldsymbol{C}(\boldsymbol{C}+\boldsymbol{D})^{-1}\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{C}+\boldsymbol{D})^{-1}\left[(\boldsymbol{C}+\boldsymbol{D}) - \boldsymbol{C}\right] = \boldsymbol{C}(\boldsymbol{C}+\boldsymbol{D})^{-1}\boldsymbol{D}.$$

The matrix in the last quadratic form equals $\boldsymbol{H}$ because

$$\left[\boldsymbol{C}(\boldsymbol{C}+\boldsymbol{D})^{-1}\boldsymbol{D}\right]^{-1} = \boldsymbol{D}^{-1}(\boldsymbol{C}+\boldsymbol{D})\boldsymbol{C}^{-1} = \boldsymbol{D}^{-1} + \boldsymbol{C}^{-1},$$

which proves the identity. □

Proof of Lemma 2. Fix a phase s whose graph update is executed and two users $i, j \in [N]$. We prove the stated implication by contraposition. Suppose the two confidence ellipsoids intersect, and choose any

$$\boldsymbol{z} \in \mathcal{E}_{i,\mathsf{e},t_s} \cap \mathcal{E}_{j,\mathsf{e},t_s}.$$

Membership in the two ellipsoids means

$$\left\|\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s} - \boldsymbol{z}\right\|^2_{\boldsymbol{V}_{i,\mathsf{e},t_s}} \le \beta^2_{i,\mathsf{e},t_s}, \qquad \left\|\widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s} - \boldsymbol{z}\right\|^2_{\boldsymbol{V}_{j,\mathsf{e},t_s}} \le \beta^2_{j,\mathsf{e},t_s}.$$

By Lemma 1 and $\lambda > 0$, both $\boldsymbol{V}_{i,\mathsf{e},t_s}$ and $\boldsymbol{V}_{j,\mathsf{e},t_s}$ are symmetric positive definite. Apply Lemma 4 with $\boldsymbol{u} = \widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s}$, $\boldsymbol{v} = \widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s}$, $\boldsymbol{C} = \boldsymbol{V}_{i,\mathsf{e},t_s}$, and $\boldsymbol{D} = \boldsymbol{V}_{j,\mathsf{e},t_s}$. By definition,

$$\boldsymbol{H}_{ij,t_s} = \left(\boldsymbol{V}^{-1}_{i,\mathsf{e},t_s} + \boldsymbol{V}^{-1}_{j,\mathsf{e},t_s}\right)^{-1}.$$

Thus, Lemma 4 yields

$$\left\|\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s} - \widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s}\right\|^2_{\boldsymbol{H}_{ij,t_s}} = \min_{\boldsymbol{w}\in\mathbb{R}^d}\left\{\left\|\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s} - \boldsymbol{w}\right\|^2_{\boldsymbol{V}_{i,\mathsf{e},t_s}} + \left\|\widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s} - \boldsymbol{w}\right\|^2_{\boldsymbol{V}_{j,\mathsf{e},t_s}}\right\}.$$

The common point $\boldsymbol{z}$ is one admissible choice of $\boldsymbol{w}$ in this minimization. Therefore, the minimum is no larger than the objective evaluated at $\boldsymbol{w} = \boldsymbol{z}$, giving

$$\begin{aligned}
&\left\|\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s} - \widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s}\right\|^2_{\boldsymbol{H}_{ij,t_s}} \\
&\le \left\|\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s} - \boldsymbol{z}\right\|^2_{\boldsymbol{V}_{i,\mathsf{e},t_s}} + \left\|\widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s} - \boldsymbol{z}\right\|^2_{\boldsymbol{V}_{j,\mathsf{e},t_s}} \\
&\le \beta^2_{i,\mathsf{e},t_s} + \beta^2_{j,\mathsf{e},t_s},
\end{aligned}$$

where the last inequality follows from $\boldsymbol{z} \in \mathcal{E}_{i,\mathsf{e},t_s} \cap \mathcal{E}_{j,\mathsf{e},t_s}$. Hence intersecting ellipsoids cannot satisfy the strict inequality in Eq. (4), which proves the contrapositive of the stated implication. □

Lemma 5 (Characterization of observable equivalence)**.** *Suppose $B \ge 1$. For every $\boldsymbol{v} \in \mathbb{R}^d$, the following statements are equivalent:*

(i) $\boldsymbol{x}_k^\top \boldsymbol{v} = 0$ *for every* $k \in [K]$.

(ii) $\Pi_W \boldsymbol{v} = \boldsymbol{0}_d$.

(iii) $\|\boldsymbol{v}\|_{\boldsymbol{\Sigma}_\mathcal{B}} = 0$.

Consequently, $\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}_\mathcal{B}} > 0$ for every nonzero $\boldsymbol{w} \in W$, that is, $\boldsymbol{\Sigma}_\mathcal{B}$ is positive definite when restricted to W, and for all users $i, j \in [N]$,

$$i \sim_{\mathrm{obs}} j \iff \Pi_W \boldsymbol{\theta}^\star_i = \Pi_W \boldsymbol{\theta}^\star_j \iff \left\|\boldsymbol{\theta}^\star_i - \boldsymbol{\theta}^\star_j\right\|_{\boldsymbol{\Sigma}_\mathcal{B}} = 0.$$

Proof of Lemma 5. Since $\mathcal{A}$ spans W, condition (i) holds if and only if $\boldsymbol{v} \in W^\perp$, which is equivalent to condition (ii). By the definition of $\boldsymbol{\Sigma}_\mathcal{B}$,

$$\|\boldsymbol{v}\|^2_{\boldsymbol{\Sigma}_\mathcal{B}} = \frac{1}{B}\sum_{\ell=1}^{B}\left(\boldsymbol{x}_{b_\ell}^\top \boldsymbol{v}\right)^2.$$

Since every term $(\boldsymbol{x}_{b_\ell}^\top \boldsymbol{v})^2$ is nonnegative, condition (iii) holds if and only if $\boldsymbol{x}_{b_\ell}^\top \boldsymbol{v} = 0$ for all $\ell \in [B]$. Since $\mathcal{B}$ also spans W, the latter condition, $\boldsymbol{x}_{b_\ell}^\top \boldsymbol{v} = 0$, is equivalent to $\boldsymbol{v} \in W^\perp$, that is, to condition (ii).

For a nonzero $\boldsymbol{w} \in W$, we have $\Pi_W \boldsymbol{w} = \boldsymbol{w} \neq \mathbf{0}_d$, so condition (ii) fails and hence $\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}_\mathcal{B}} > 0$. Finally, by Definition 1, $i \sim_{\text{obs}} j$ means that condition (i) holds for $\boldsymbol{v} = \boldsymbol{\theta}_i^\star - \boldsymbol{\theta}_j^\star$, and the equivalence of conditions (i), (ii), and (iii) for such $\boldsymbol{v}$ gives

$$i \sim_{\text{obs}} j \iff \Pi_W \boldsymbol{\theta}_i^\star = \Pi_W \boldsymbol{\theta}_j^\star \iff \left\|\boldsymbol{\theta}_i^\star - \boldsymbol{\theta}_j^\star\right\|_{\boldsymbol{\Sigma}_\mathcal{B}} = 0.$$

□

Lemma 6 (Confidence for block-committed exploration estimators)**.** *Under Assumptions 1 and 2, consider the block-committed exploration estimators with $\lambda \geq 1$ and $\delta \in (0,1)$. For every user and round, write*

$$\beta_{i,\mathsf{e},t} \coloneqq \sqrt{2\log\frac{N}{\delta} + d\log\left(1 + \frac{n_{i,\mathsf{e},t}}{\lambda d}\right)} + \sqrt{\lambda}.$$

Define

$$\mathcal{E}_\mathsf{e} \coloneqq \left\{ \left\| \widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t} - \Pi_W \boldsymbol{\theta}_i^\star \right\|_{\boldsymbol{V}_{i,\mathsf{e},t}} \leq \beta_{i,\mathsf{e},t} \; \forall\, i \in [N],\; 0 \leq t \leq T \right\}.$$

Then $\mathbb{P}(\mathcal{E}_\mathsf{e}) \geq 1 - \delta$.

Proof of Lemma 6. Fix a user $i \in [N]$. We first construct a predictable process using all of this user's raw forced-exploration observations, before restricting attention to completed blocks. Let $I_\mathsf{e}(\tau)$ indicate whether round τ belongs to a forced-exploration subphase, and define

$$\boldsymbol{z}_{i,\tau} \coloneqq I_\mathsf{e}(\tau)\mathbf{1}\{u_\tau = i\}\boldsymbol{x}_{a_\tau}.$$

Thus, $\boldsymbol{z}_{i,\tau}$ equals the arm feature selected at round τ, exactly when that round is a forced-exploration round serving user i, and equals $\mathbf{0}_d$ otherwise. The phase schedule is deterministic, and in every forced-exploration round, the selected arm a_τ is determined by the pointer of the served user u_τ, which depends only on the preceding rounds. Hence $\boldsymbol{z}_{i,\tau}$ is $\mathcal{F}_{\tau-1}$-measurable.

Define the auxiliary raw exploration statistics, which aggregate all forced-exploration observations of user i up to round t before applying the block-commit rule, as

$$\boldsymbol{V}_{i,\text{raw},t} \coloneqq \lambda \boldsymbol{I}_d + \sum_{\tau=1}^{t} \boldsymbol{z}_{i,\tau}\boldsymbol{z}_{i,\tau}^\top, \qquad \boldsymbol{b}_{i,\text{raw},t} \coloneqq \sum_{\tau=1}^{t} \boldsymbol{z}_{i,\tau} y_\tau,$$

and let $\widehat{\boldsymbol{\theta}}_{i,\text{raw},t} \coloneqq \boldsymbol{V}_{i,\text{raw},t}^{-1}\boldsymbol{b}_{i,\text{raw},t}$. Whenever $\boldsymbol{z}_{i,\tau} \neq \mathbf{0}_d$, the current user is i, and projection onto W preserves the selected arm's mean reward; hence, for every τ,

$$\boldsymbol{z}_{i,\tau} y_\tau = \boldsymbol{z}_{i,\tau}\boldsymbol{z}_{i,\tau}^\top \Pi_W \boldsymbol{\theta}_i^\star + \boldsymbol{z}_{i,\tau}\epsilon_\tau,$$

where both sides vanish when $\boldsymbol{z}_{i,\tau} = \mathbf{0}_d$. Summing over $\tau \leq t$ yields

$$\boldsymbol{b}_{i,\text{raw},t} = \left(\boldsymbol{V}_{i,\text{raw},t} - \lambda \boldsymbol{I}_d\right)\Pi_W \boldsymbol{\theta}_i^\star + \sum_{\tau=1}^{t} \boldsymbol{z}_{i,\tau}\epsilon_\tau.$$

Moreover, $\|\Pi_W \boldsymbol{\theta}_i^\star\| \leq \|\boldsymbol{\theta}_i^\star\| \leq 1$ by Assumption 2. Applying the self-normalized concentration inequality (Abbasi-yadkori et al., 2011) with failure probability δ/N, we obtain, with probability at least $1 - \delta/N$, for every $0 \leq t \leq T$,

$$\left\| \widehat{\boldsymbol{\theta}}_{i,\text{raw},t} - \Pi_W \boldsymbol{\theta}_i^\star \right\|_{\boldsymbol{V}_{i,\text{raw},t}} \leq \sqrt{2\log\frac{N}{\delta} + \log\frac{\det(\boldsymbol{V}_{i,\text{raw},t})}{\det(\lambda \boldsymbol{I}_d)}} + \sqrt{\lambda}.$$

We now relate the auxiliary raw statistics to the block-committed exploration statistics. Let $\tau_{i,\ell}$ denote the round of user i's ℓth raw forced-exploration observation and set $\tau_{i,0} = 0$. Fix any

round $t \in \{0, \dots, T\}$. By definition, the committed statistics consist exactly of the first $n_{i,\mathsf{e},t}$ raw forced-exploration observations of user i, so

$$\boldsymbol{V}_{i,\mathsf{e},t} = \boldsymbol{V}_{i,\mathrm{raw},\tau_{i,n_{i,\mathsf{e},t}}}, \qquad \boldsymbol{b}_{i,\mathsf{e},t} = \boldsymbol{b}_{i,\mathrm{raw},\tau_{i,n_{i,\mathsf{e},t}}},$$

and hence $\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t} = \widehat{\boldsymbol{\theta}}_{i,\mathrm{raw},\tau_{i,n_{i,\mathsf{e},t}}}$. Since $\tau_{i,n_{i,\mathsf{e},t}} \le t \le T$ and the confidence bound holds simultaneously for every round in $\{0, \dots, T\}$, the bound also holds at the possibly random round $\tau_{i,n_{i,\mathsf{e},t}}$, which gives

$$\left\|\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t} - \Pi_W \boldsymbol{\theta}_i^\star\right\|_{\boldsymbol{V}_{i,\mathsf{e},t}} \le \sqrt{2\log\frac{N}{\delta} + \log\frac{\det(\boldsymbol{V}_{i,\mathsf{e},t})}{\det(\lambda \boldsymbol{I}_d)}} + \sqrt{\lambda}.$$

The committed design matrix contains $n_{i,\mathsf{e},t}$ arm observations, each with norm at most one, and hence

$$\mathrm{tr}\left(\boldsymbol{V}_{i,\mathsf{e},t} - \lambda \boldsymbol{I}_d\right) = \sum_{\ell=1}^{n_{i,\mathsf{e},t}} \left\|\boldsymbol{x}_{a_{\tau_{i,\ell}}}\right\|^2 \le n_{i,\mathsf{e},t}.$$

Using the determinant-trace inequality,

$$\log\frac{\det(\boldsymbol{V}_{i,\mathsf{e},t})}{\det(\lambda \boldsymbol{I}_d)} \le d\log\left(1 + \frac{\mathrm{tr}\left(\boldsymbol{V}_{i,\mathsf{e},t} - \lambda \boldsymbol{I}_d\right)}{\lambda d}\right) \le d\log\left(1 + \frac{n_{i,\mathsf{e},t}}{\lambda d}\right).$$

Therefore,

$$\left\|\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t} - \Pi_W \boldsymbol{\theta}_i^\star\right\|_{\boldsymbol{V}_{i,\mathsf{e},t}} \le \sqrt{2\log\frac{N}{\delta} + d\log\left(1 + \frac{n_{i,\mathsf{e},t}}{\lambda d}\right)} + \sqrt{\lambda},$$

whose right-hand side equals $\beta_{i,\mathsf{e},t}$ by definition. Finally, taking a union bound over the N users gives total failure probability at most $N(\delta/N) = \delta$, proving $\mathbb{P}(\mathcal{E}_\mathsf{e}) \ge 1 - \delta$. □

Proof of Corollary 1. Work on the event $\mathcal{E}_\mathsf{e}$ from Lemma 6, which has probability at least $1 - \delta$. Fix a phase s whose graph update is executed and two users i, j in the same observable cluster $\mathcal{I}_c$. By definition of $\mathcal{I}_c$, it holds that

$$\Pi_W \boldsymbol{\theta}_i^\star = \Pi_W \boldsymbol{\theta}_j^\star = \boldsymbol{\vartheta}_c^\star.$$

On $\mathcal{E}_\mathsf{e}$,

$$\left\|\boldsymbol{\vartheta}_c^\star - \widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s}\right\|_{\boldsymbol{V}_{i,\mathsf{e},t_s}} \le \beta_{i,\mathsf{e},t_s}, \qquad \left\|\boldsymbol{\vartheta}_c^\star - \widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s}\right\|_{\boldsymbol{V}_{j,\mathsf{e},t_s}} \le \beta_{j,\mathsf{e},t_s},$$

thus,

$$\boldsymbol{\vartheta}_c^\star \in \mathcal{E}_{i,\mathsf{e},t_s} \cap \mathcal{E}_{j,\mathsf{e},t_s}.$$

Since the two ellipsoids intersect, Lemma 2 implies that Eq. (4) fails for (i, j), and hence the edge (i, j) is not deleted at the graph update of phase s. Because $\mathcal{E}_\mathsf{e}$ controls all users and all rounds simultaneously, the same argument applies to every within-observable-cluster pair at every executed graph update. Since the graph starts complete and edges are deleted only at graph updates, every edge between users in the same observable cluster remains present throughout the entire horizon. □

C.4 Proof of Theorem 1

Lemma 7 (Precision of surviving edges). *Let $\beta \coloneqq \sqrt{2\log(N/\delta) + d\log(1 + T/(\lambda d))} + \sqrt{\lambda}$ denote the confidence coefficient used in Algorithm 1. Suppose $\mathcal{E}_\mathsf{e}$ holds and the edge (i, j) survives the executed graph update of phase s. If $n_{\min} \coloneqq \min\{n_{i,\mathsf{e},t_s}, n_{j,\mathsf{e},t_s}\} > 0$, then*

$$\left\|\boldsymbol{\vartheta}_{m(i)}^\star - \boldsymbol{\vartheta}_{m(j)}^\star\right\|_{\boldsymbol{\Sigma}_\mathcal{B}} \le \frac{(2 + 2\sqrt{2})\beta}{\sqrt{n_{\min}}}.$$

Proof of Lemma 7. By Lemma 1, $\boldsymbol{V}_{u,\mathsf{e},t_s} = \lambda \boldsymbol{I}_d + n_{u,\mathsf{e},t_s} \boldsymbol{\Sigma}_\mathcal{B}$ for every user $u \in [N]$. Since $n_{i,\mathsf{e},t_s} \ge n_{\min}$ by the definition of $n_{\min}$ and $\boldsymbol{\Sigma}_\mathcal{B} \succeq \boldsymbol{0}_{d\times d}$,

$$\boldsymbol{V}_{i,\mathsf{e},t_s} - \left(\lambda \boldsymbol{I}_d + n_{\min} \boldsymbol{\Sigma}_\mathcal{B}\right) = \left(n_{i,\mathsf{e},t_s} - n_{\min}\right) \boldsymbol{\Sigma}_\mathcal{B} \succeq \boldsymbol{0}_{d\times d},$$

and the same argument applies to user j. Therefore, $\boldsymbol{V}_{i,\mathsf{e},t_s} \succeq \lambda \boldsymbol{I}_d + n_{\min}\boldsymbol{\Sigma}_{\mathcal{B}}$ and $\boldsymbol{V}_{j,\mathsf{e},t_s} \succeq \lambda \boldsymbol{I}_d + n_{\min}\boldsymbol{\Sigma}_{\mathcal{B}}$. Because $\lambda > 0$, the matrix $\lambda \boldsymbol{I}_d + n_{\min}\boldsymbol{\Sigma}_{\mathcal{B}}$ is positive definite, and so are $\boldsymbol{V}_{i,\mathsf{e},t_s}$ and $\boldsymbol{V}_{j,\mathsf{e},t_s}$. Since matrix inversion reverses the Loewner order on positive-definite matrices,

$$\boldsymbol{V}_{i,\mathsf{e},t_s}^{-1} \preceq (\lambda \boldsymbol{I}_d + n_{\min}\boldsymbol{\Sigma}_{\mathcal{B}})^{-1}, \qquad \boldsymbol{V}_{j,\mathsf{e},t_s}^{-1} \preceq (\lambda \boldsymbol{I}_d + n_{\min}\boldsymbol{\Sigma}_{\mathcal{B}})^{-1}.$$

Adding the two inequalities gives

$$\boldsymbol{V}_{i,\mathsf{e},t_s}^{-1} + \boldsymbol{V}_{j,\mathsf{e},t_s}^{-1} \preceq 2\,(\lambda \boldsymbol{I}_d + n_{\min}\boldsymbol{\Sigma}_{\mathcal{B}})^{-1}.$$

Applying order reversal once more to $\boldsymbol{H}_{ij,t_s} = (\boldsymbol{V}_{i,\mathsf{e},t_s}^{-1} + \boldsymbol{V}_{j,\mathsf{e},t_s}^{-1})^{-1}$ yields

$$\begin{aligned}\boldsymbol{H}_{ij,t_s} &\succeq \frac{1}{2}\,(\lambda \boldsymbol{I}_d + n_{\min}\boldsymbol{\Sigma}_{\mathcal{B}}) \\ &= \frac{\lambda}{2}\boldsymbol{I}_d + \frac{n_{\min}}{2}\boldsymbol{\Sigma}_{\mathcal{B}} \\ &\succeq \frac{n_{\min}}{2}\boldsymbol{\Sigma}_{\mathcal{B}}. \end{aligned} \tag{9}$$

Moreover, since $\boldsymbol{V}_{j,\mathsf{e},t_s}^{-1}$ is positive definite, $\boldsymbol{V}_{i,\mathsf{e},t_s}^{-1} + \boldsymbol{V}_{j,\mathsf{e},t_s}^{-1} \succeq \boldsymbol{V}_{i,\mathsf{e},t_s}^{-1}$, and order reversal gives $\boldsymbol{H}_{ij,t_s} \preceq \boldsymbol{V}_{i,\mathsf{e},t_s}$; similarly, $\boldsymbol{H}_{ij,t_s} \preceq \boldsymbol{V}_{j,\mathsf{e},t_s}$.

Because the edge (i,j) survives the graph update of phase s, Eq. (4) is not satisfied, that is,

$$\left\|\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s} - \widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s}\right\|_{\boldsymbol{H}_{ij,t_s}} \le \sqrt{\beta_{i,\mathsf{e},t_s}^2 + \beta_{j,\mathsf{e},t_s}^2}.$$

On the event $\mathcal{E}_{\mathsf{e}}$,

$$\left\|\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s} - \Pi_W\boldsymbol{\theta}_i^\star\right\|_{\boldsymbol{V}_{i,\mathsf{e},t_s}} \le \beta_{i,\mathsf{e},t_s}.$$

Since $\boldsymbol{H}_{ij,t_s} \preceq \boldsymbol{V}_{i,\mathsf{e},t_s}$, the corresponding quadratic form is no larger, and therefore

$$\left\|\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s} - \Pi_W\boldsymbol{\theta}_i^\star\right\|_{\boldsymbol{H}_{ij,t_s}} \le \beta_{i,\mathsf{e},t_s}.$$

Likewise,

$$\left\|\widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s} - \Pi_W\boldsymbol{\theta}_j^\star\right\|_{\boldsymbol{H}_{ij,t_s}} \le \beta_{j,\mathsf{e},t_s}.$$

Applying the triangle inequality to

$$\Pi_W\boldsymbol{\theta}_i^\star - \Pi_W\boldsymbol{\theta}_j^\star = \left(\Pi_W\boldsymbol{\theta}_i^\star - \widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s}\right) + \left(\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s} - \widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s}\right) + \left(\widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s} - \Pi_W\boldsymbol{\theta}_j^\star\right),$$

we obtain

$$\begin{aligned}\left\|\Pi_W\boldsymbol{\theta}_i^\star - \Pi_W\boldsymbol{\theta}_j^\star\right\|_{\boldsymbol{H}_{ij,t_s}} &\le \left\|\Pi_W\boldsymbol{\theta}_i^\star - \widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s}\right\|_{\boldsymbol{H}_{ij,t_s}} + \left\|\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s} - \widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s}\right\|_{\boldsymbol{H}_{ij,t_s}} + \left\|\widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s} - \Pi_W\boldsymbol{\theta}_j^\star\right\|_{\boldsymbol{H}_{ij,t_s}} \\ &\le \beta_{i,\mathsf{e},t_s} + \sqrt{\beta_{i,\mathsf{e},t_s}^2 + \beta_{j,\mathsf{e},t_s}^2} + \beta_{j,\mathsf{e},t_s} \\ &\le (2+\sqrt{2})\beta,\end{aligned}$$

where the last inequality follows from $\beta_{i,\mathsf{e},t_s}, \beta_{j,\mathsf{e},t_s} \le \beta$, which holds because $n_{i,\mathsf{e},t_s}, n_{j,\mathsf{e},t_s} \le T$.

We now convert this bound into the $\boldsymbol{\Sigma}_{\mathcal{B}}$-norm. By Eq. (9), every $\boldsymbol{u} \in \mathbb{R}^d$ satisfies

$$\boldsymbol{u}^\top \boldsymbol{H}_{ij,t_s}\boldsymbol{u} \ge \frac{n_{\min}}{2}\boldsymbol{u}^\top\boldsymbol{\Sigma}_{\mathcal{B}}\boldsymbol{u},$$

equivalently,

$$\|\boldsymbol{u}\|_{\boldsymbol{\Sigma}_{\mathcal{B}}} \le \sqrt{\frac{2}{n_{\min}}}\|\boldsymbol{u}\|_{\boldsymbol{H}_{ij,t_s}}.$$

Applying this inequality with $\boldsymbol{u} = \Pi_W\boldsymbol{\theta}_i^\star - \Pi_W\boldsymbol{\theta}_j^\star$ gives

$$\begin{aligned}\left\|\Pi_W\boldsymbol{\theta}_i^\star - \Pi_W\boldsymbol{\theta}_j^\star\right\|_{\boldsymbol{\Sigma}_{\mathcal{B}}} &\le \sqrt{\frac{2}{n_{\min}}}\left\|\Pi_W\boldsymbol{\theta}_i^\star - \Pi_W\boldsymbol{\theta}_j^\star\right\|_{\boldsymbol{H}_{ij,t_s}} \\ &\le \frac{\sqrt{2}(2+\sqrt{2})\beta}{\sqrt{n_{\min}}} \\ &= \frac{(2+2\sqrt{2})\beta}{\sqrt{n_{\min}}}.\end{aligned}$$

Since $\Pi_W\boldsymbol{\theta}_i^\star = \boldsymbol{\vartheta}_{m(i)}^\star$ and $\Pi_W\boldsymbol{\theta}_j^\star = \boldsymbol{\vartheta}_{m(j)}^\star$, the claim follows. □

Lemma 8 (Phase count and coverage under Categorical arrivals)**.** *Under the Categorical arrival model in Eq.* (1)*, use the phase lengths in Algorithm 1, with $B \geq 1$. Let $J_T \coloneqq 1 + \lceil \log_2(T+1) \rceil$. At most J_T phases start, and at most J_T graph updates are executed, by time T. Moreover, with probability at least $1 - J_T\delta$,*

$$n_{i,\mathsf{e},t_s} \geq N\pi_{\min} 2^s n_0^\star$$

simultaneously for every user i and every phase satisfying $t_s \leq T$ and $N\pi_{\min}2^s \geq 1$.

Henceforth, we denote by $\mathcal{E}_{\mathrm{cov}}$ the event on which the coverage bound stated in Lemma 8 holds simultaneously for all $i \in [N]$ and all phases satisfying $t_s \leq T$ and $N\pi_{\min}2^s \geq 1$.

Proof of Lemma 8. Recall that $h_s = \sum_{q=0}^{s-1}(M_q + L_q)$, with $h_0 = 0$, denotes the total number of rounds scheduled before phase s. We first bound the number of phases that can start by time T. If phase s starts, then all preceding phases have completed, so its first round satisfies $h_s + 1 \leq T$. Since $L_q \geq 0$ for all $0 \leq q \leq s-1$,

$$h_s = \sum_{q=0}^{s-1}(M_q + L_q) \geq \sum_{q=0}^{s-1} M_q = C_0(2^s - 1).$$

Hence $C_0(2^s - 1) + 1 \leq h_s + 1 \leq T$, and since $C_0 \geq 1$,

$$2^s \leq 1 + \frac{T-1}{C_0} \leq T.$$

Therefore,

$$s \leq \lfloor \log_2 T \rfloor,$$

and hence

$$\begin{aligned} |\{s \in \mathbb{Z}_{\geq 0} : h_s + 1 \leq T\}| &\leq 1 + \lfloor \log_2 T \rfloor \\ &\leq 1 + \lceil \log_2(T+1) \rceil \\ &= J_T. \end{aligned}$$

Since each phase contains at most one graph update, at most J_T graph updates are executed by time T as well.

Now, fix a user i and a phase s whose forced-exploration subphase is completed by time T, and assume $N\pi_{\min}2^s \geq 1$. The $M_s = C_0 2^s$ forced-exploration rounds of this phase occur at deterministic positions in the schedule. Since user arrivals are IID, the number $X_{i,s}$ of these rounds at which user i arrives satisfies

$$X_{i,s} \sim \mathrm{Binomial}(M_s, \pi_i).$$

Its mean is

$$\begin{aligned} \mu_{i,s} \coloneqq \mathbb{E}[X_{i,s}] &= \pi_i C_0 2^s \\ &\geq \pi_{\min} C_0 2^s \\ &\geq N\pi_{\min} 2^s \left(16 \log \frac{N}{\delta} + 4n_0^\star\right), \end{aligned}$$

where the last line uses $C_0 \geq 16N\log(N/\delta) + 4Nn_0^\star$.

For completeness, we derive the lower-tail estimate used below. For any $\alpha > 0$, Markov's inequality applied to $\exp(-\alpha X_{i,s})$ gives

$$\mathbb{P}\left(X_{i,s} < \frac{\mu_{i,s}}{2}\right) \leq \exp\left(\frac{\alpha\mu_{i,s}}{2}\right) \mathbb{E}\left[\exp(-\alpha X_{i,s})\right].$$

Choose $\alpha = \log 2$. The moment-generating function of a binomial variable gives

$$\mathbb{E}[\exp(-(\log 2)X_{i,s})] = (1 - \pi_i + \pi_i/2)^{M_s} = (1 - \pi_i/2)^{M_s}.$$

Using $1 - x \le \exp(-x)$, we obtain

$$\begin{aligned}\mathbb{P}\left(X_{i,s} < \frac{\mu_{i,s}}{2}\right) &\le \exp\left(\frac{\log 2}{2}\mu_{i,s} - \frac{1}{2}M_s\pi_i\right) \\ &= \exp\left(-\frac{1-\log 2}{2}\mu_{i,s}\right) \\ &\le \exp\left(-\frac{\mu_{i,s}}{8}\right),\end{aligned}$$

where $1 - \log 2 > 1/4$ is used in the last step. Because $N\pi_{\min}2^s \ge 1$, the preceding lower bound on $\mu_{i,s}$ implies

$$\begin{aligned}\exp(-\mu_{i,s}/8) &\le \exp\left(-2\log\frac{N}{\delta}\right) \\ &= \left(\frac{\delta}{N}\right)^2 \\ &\le \frac{\delta}{N}.\end{aligned}$$

Therefore, with probability at least $1 - \delta/N$,

$$X_{i,s} \ge \frac{1}{2}\mu_{i,s} \ge 2N\pi_{\min}2^s n_0^\star.$$

Recall that $E_i(t_s)$ denotes the number of forced-exploration observations of user i through round t_s, the last round of the phase-s exploration subphase. The count $E_i(t_s)$ includes all observations counted by $X_{i,s}$ and possibly observations from earlier phases, so $E_i(t_s) \ge X_{i,s}$. As noted in Appendix C.2, $n_{i,\mathsf{e},t_s} = B\lfloor E_i(t_s)/B \rfloor$, since at most $B-1$ observations remain in the incomplete buffer; hence

$$\begin{aligned}n_{i,\mathsf{e},t_s} &= B\left\lfloor \frac{E_i(t_s)}{B} \right\rfloor \\ &\ge E_i(t_s) - B \\ &\ge 2N\pi_{\min}2^s n_0^\star - B.\end{aligned}$$

Since $n_0^\star = B\lceil (2+2\sqrt{2})^2\beta^2/B \rceil \ge B$ and $N\pi_{\min}2^s \ge 1$,

$$2N\pi_{\min}2^s n_0^\star - B \ge N\pi_{\min}2^s n_0^\star.$$

Thus the desired committed-count lower bound holds for this user and phase except on an event of probability at most δ/N. A union bound over the N users costs at most δ for each completed exploration subphase, and a second union bound over at most J_T such phases gives total failure probability at most $J_T\delta$. □

Proof of Theorem 1. Let $\mathcal{E}_{\mathsf{e}}$ and $\mathcal{E}_{\mathrm{cov}}$ denote the events defined in Lemma 6 and after Lemma 8, respectively. By these lemmas, $\mathbb{P}(\mathcal{E}_{\mathsf{e}}) \ge 1 - \delta$ and $\mathbb{P}(\mathcal{E}_{\mathrm{cov}}) \ge 1 - J_T\delta$. A union bound therefore gives

$$\mathbb{P}(\mathcal{E}_{\mathsf{e}} \cap \mathcal{E}_{\mathrm{cov}}) \ge 1 - (J_T + 1)\delta.$$

We establish all claims on this intersection.

First, by Corollary 1, no edge between users in the same observable cluster is ever deleted: on $\mathcal{E}_{\mathsf{e}}$, the confidence ellipsoids of two such users contain the same observable parameter and therefore intersect, regardless of how many exploration samples have been committed. Thus, this part of the argument does not require the coverage event $\mathcal{E}_{\mathrm{cov}}$.

If $C = 1$, every pair of users is a within-cluster pair, so the graph remains complete throughout the execution. Statement (i) then holds trivially because its left-hand side is zero, and statement (ii) holds with $s_\pi := 0$ because $\mathcal{N}_s(i) = [N] = \mathcal{I}_1$ for every $i \in [N]$ and every executed update. Hence, we assume $C \ge 2$ henceforth.

Next, consider a phase s whose graph update is executed and satisfies $N\pi_{\min}2^s \ge 1$, and let $(i,j) \in E(\mathcal{G}_s)$ be any surviving edge. Since the update is executed, $t_s + 1 \le T$, so on $\mathcal{E}_{\mathrm{cov}}$,

$$\min\{n_{i,\mathsf{e},t_s}, n_{j,\mathsf{e},t_s}\} \ge N\pi_{\min}2^s n_0^\star > 0.$$

Applying Lemma 7 gives

$$\begin{aligned}\left\|\boldsymbol{\vartheta}^\star_{m(i)} - \boldsymbol{\vartheta}^\star_{m(j)}\right\|_{\boldsymbol{\Sigma}_\mathcal{B}} &\le \frac{(2+2\sqrt{2})\beta}{\sqrt{\min\{n_{i,\mathsf{e},t_s}, n_{j,\mathsf{e},t_s}\}}} \\ &\le \frac{(2+2\sqrt{2})\beta}{\sqrt{N\pi_{\min}2^s n_0^\star}} \\ &\le \frac{2^{-s/2}}{\sqrt{N\pi_{\min}}} \\ &= \frac{\rho_s}{\sqrt{N\pi_{\min}}},\end{aligned}$$

where the last inequality uses $n_0^\star \ge (2+2\sqrt{2})^2\beta^2$, which follows from $n_0^\star = B\lceil(2+2\sqrt{2})^2\beta^2/B\rceil$. The preceding display proves statement (i).

Now recall that

$$s_\pi \coloneqq \min\left\{s \in \mathbb{Z}_{\ge 0} : N\pi_{\min}2^s \ge 1 \text{ and } \frac{\rho_s}{\sqrt{N\pi_{\min}}} < \Delta_\mathcal{B}\right\}. \tag{10}$$

Both defining conditions are lower bounds on 2^s: the first reads $2^s \ge 1/(N\pi_{\min})$, and the second is equivalent to $2^s > 1/(N\pi_{\min}\Delta_\mathcal{B}^2)$. Since $N\pi_{\min} > 0$ and $\Delta_\mathcal{B} > 0$, both thresholds are finite, so the set in Eq. (10) is nonempty and s_π is well defined; moreover, since 2^s increases with s, both conditions continue to hold for every $s \ge s_\pi$.

If a cross-cluster edge (i, j) survived the executed graph update of a phase $s \ge s_\pi$, then the definition of $\Delta_\mathcal{B}$ and statement (i) would give

$$\Delta_\mathcal{B} \le \left\|\boldsymbol{\vartheta}^\star_{m(i)} - \boldsymbol{\vartheta}^\star_{m(j)}\right\|_{\boldsymbol{\Sigma}_\mathcal{B}} \le \frac{\rho_s}{\sqrt{N\pi_{\min}}} < \Delta_\mathcal{B},$$

which is a contradiction. Thus no cross-cluster edge remains after any executed update with $s \ge s_\pi$. Since every within-cluster edge is preserved, the graph is a disjoint union of complete subgraphs on the observable clusters. Therefore, each self-inclusive direct neighborhood is exactly

$$\mathcal{N}_s(i) = \mathcal{I}_{m(i)},$$

for all $i \in [N]$, which proves statement (ii). □

C.5 Proof of Theorem 2

Lemma 9 (Exploration and pre-recovery duration bounds)**.** *Use the phase schedule and calibration in Algorithm 1 with $B \ge 1$, and the recovery index in Eq.* (10)*, setting $s_\pi = 0$ when $C = 1$. Let $I_\mathsf{e}(t)$ indicate a forced-exploration round, let $E(T) \coloneqq \sum_{t=1}^T I_\mathsf{e}(t)$ be the number of forced-exploration rounds executed by time T, and define*

$$D_\pi \coloneqq \sum_{s=0}^{s_\pi - 1} L_s,$$

with an empty sum equal to zero. Then, deterministically,

$$E(T) \le 4C_0^p T^{1-p}.$$

For $C \ge 2$,

$$D_\pi \le 4C_0\left(\frac{\max\{1, \Delta_\mathcal{B}^{-2}\}}{N\pi_{\min}}\right)^{\frac{1}{1-p}} \le 4C_0\left(\frac{4}{N\pi_{\min}\Delta_\mathcal{B}^2}\right)^{\frac{1}{1-p}}. \tag{11}$$

If $s_\pi = 0$, including the case $C = 1$, then $D_\pi = 0$. Write $r_t \coloneqq (\boldsymbol{x}_{a_t^\star} - \boldsymbol{x}_{a_t})^\top \boldsymbol{\vartheta}^\star_{m(u_t)}$ for the instantaneous regret. Under Assumption 2, $0 \le r_t \le 2$, and hence

$$\mathrm{Reg}_{\mathrm{explore}}(T) \le 8C_0^p T^{1-p}, \qquad \mathrm{Reg}_{\mathrm{pre}}(T) \le 2D_\pi,$$

where $\mathrm{Reg}_{\mathrm{pre}}(T)$ accumulates the regret of adaptive-selection rounds in phases $s < s_\pi$.

Proof of Lemma 9. We first bound the number of executed forced-exploration rounds. Let $\bar{s}$ be the last phase that starts by time T. If $\bar{s} = 0$, every forced-exploration round executed by time T belongs to phase 0, whose forced-exploration subphase has length $M_0 = C_0$. Therefore,

$$E(T) \le \min\{T, C_0\}.$$

For any $a, b > 0$ and $p \in (0,1)$, $\min\{a,b\} \le a^{1-p}b^p$. Applying this inequality with $a = T$ and $b = C_0$ gives

$$E(T) \le C_0^p T^{1-p}.$$

Now suppose $\bar{s} \ge 1$. For phase $\bar{s}$ to start, phase $\bar{s}-1$ must have been completed, including its adaptive-selection subphase. Hence

$$T \ge L_{\bar{s}-1} \ge C_0 2^{\frac{\bar{s}-1}{1-p}}.$$

Rearranging gives

$$2^{\bar{s}-1} \le \left(\frac{T}{C_0}\right)^{1-p}.$$

The number of executed exploration rounds is no larger than the total scheduled exploration length through phase $\bar{s}$, even if the horizon ends partway through that phase. Therefore,

$$\begin{aligned} E(T) &\le \sum_{s=0}^{\bar{s}} M_s \\ &= C_0 \sum_{s=0}^{\bar{s}} 2^s \\ &= C_0(2^{\bar{s}+1} - 1) \\ &< 4C_0 2^{\bar{s}-1} \\ &\le 4C_0 \left(\frac{T}{C_0}\right)^{1-p} \\ &= 4C_0^p T^{1-p}. \end{aligned}$$

The two cases together prove the exploration-duration bound.

We next bound the scheduled adaptive duration before recovery. If $s_\pi = 0$, the sum defining D_π is empty and the claim is immediate. Assume henceforth that $s_\pi \ge 1$, which also implies $C \ge 2$. By minimality of s_π, phase $s_\pi - 1$ fails at least one of the two conditions in Eq. (10):

$$N\pi_{\min} 2^{s_\pi - 1} < 1 \qquad \text{or} \qquad \frac{\rho_{s_\pi - 1}}{\sqrt{N\pi_{\min}}} \ge \Delta_{\mathcal{B}}.$$

The first alternative yields

$$2^{s_\pi - 1} < \frac{1}{N\pi_{\min}},$$

whereas squaring and rearranging the second, with $\rho_{s_\pi - 1} = 2^{-(s_\pi - 1)/2}$, yields

$$2^{s_\pi - 1} \le \frac{1}{N\pi_{\min}\Delta_{\mathcal{B}}^2}.$$

Combining the two possibilities gives

$$2^{s_\pi - 1} \le \frac{1}{N\pi_{\min}} \max\{1, \Delta_{\mathcal{B}}^{-2}\} \le \frac{4}{N\pi_{\min}\Delta_{\mathcal{B}}^2}, \tag{12}$$

where the final inequality holds because $0 < \Delta_{\mathcal{B}} \le 2$ (see Section 3.2) implies both $1 \le 4\Delta_{\mathcal{B}}^{-2}$ and $\Delta_{\mathcal{B}}^{-2} \le 4\Delta_{\mathcal{B}}^{-2}$.

Since $C_0 \ge 1$ and $2^{s/(1-p)} \ge 1$, the elementary inequality $\lceil x \rceil \le 2x$ for $x \ge 1$ gives

$$L_s \le 2C_0 2^{\frac{s}{1-p}}.$$

Therefore,

$$D_\pi = \sum_{s=0}^{s_\pi - 1} L_s \le 2C_0 \sum_{s=0}^{s_\pi - 1} 2^{\frac{s}{1-p}}.$$

Factoring out the largest term by setting $h = s_\pi - 1 - s$ gives

$$\begin{aligned} D_\pi &\le 2C_0 2^{\frac{s_\pi - 1}{1-p}} \sum_{h=0}^{s_\pi - 1} 2^{-\frac{h}{1-p}} \\ &\le \frac{2C_0}{1 - 2^{-\frac{1}{1-p}}} 2^{\frac{s_\pi - 1}{1-p}} \\ &\le 4C_0 2^{\frac{s_\pi - 1}{1-p}}, \end{aligned}$$

where the second line extends the finite sum to an infinite geometric series and the last line uses $2^{-1/(1-p)} \le 1/2$. Since $x \mapsto x^{1/(1-p)}$ is increasing on $[0, \infty)$, substituting Eq. (12) proves both displayed upper bounds on D_π. If the horizon ends before all pre-recovery phases are executed, D_π only overcounts the actual number of pre-recovery adaptive rounds.

Finally, $r_t \ge 0$ by the optimality of $a_t^\star$, and $r_t \le 2$ by the Cauchy-Schwarz inequality, Assumption 2, and $\|\boldsymbol{\vartheta}^\star_{m(u_t)}\| \le 1$. Multiplying the relevant round counts by two therefore yields the two regret bounds. □

Lemma 10 (Time-uniform confidence for fixed observable clusters)**.** *Under Assumptions 1 and 2, let* $\beta \coloneqq \sqrt{2\log(N/\delta) + d\log(1 + T/(\lambda d))} + \sqrt{\lambda}$*. For each true observable cluster* $\mathcal{I}_c$*, define the statistics constructed from all observations of that cluster by*

$$\boldsymbol{V}_{\mathcal{I}_c,t} \coloneqq \lambda \boldsymbol{I}_d + \sum_{\tau=1}^{t} \mathbf{1}\{u_\tau \in \mathcal{I}_c\} \boldsymbol{x}_{a_\tau} \boldsymbol{x}_{a_\tau}^\top, \qquad \boldsymbol{b}_{\mathcal{I}_c,t} \coloneqq \sum_{\tau=1}^{t} \mathbf{1}\{u_\tau \in \mathcal{I}_c\} y_\tau \boldsymbol{x}_{a_\tau}.$$

Let $\widehat{\boldsymbol{\theta}}_{\mathcal{I}_c,t} \coloneqq \boldsymbol{V}_{\mathcal{I}_c,t}^{-1} \boldsymbol{b}_{\mathcal{I}_c,t}$ *and* $T_c(t) \coloneqq \sum_{\tau=1}^{t} \mathbf{1}\{u_\tau \in \mathcal{I}_c\}$*. Define*

$$\mathcal{E}_{\mathrm{cl}} \coloneqq \left\{ \left\| \widehat{\boldsymbol{\theta}}_{\mathcal{I}_c,t} - \boldsymbol{\vartheta}_c^\star \right\|_{\boldsymbol{V}_{\mathcal{I}_c,t}} \le \beta,\ \forall c \in [C],\ 0 \le t \le T \right\}.$$

Then $\mathbb{P}(\mathcal{E}_{\mathrm{cl}}) \ge 1 - \delta$*.*

Proof of Lemma 10. Fix an observable cluster $\mathcal{I}_c$. Although the learner does not know $\mathcal{I}_c$, the set $\mathcal{I}_c$ is a fixed subset of users determined by the problem instance. Define the masked feature

$$\boldsymbol{z}_{c,\tau} \coloneqq \mathbf{1}\{u_\tau \in \mathcal{I}_c\} \boldsymbol{x}_{a_\tau}.$$

In every round, the selected arm a_τ is a deterministic function of the preceding rounds and the arriving user u_τ: the round-robin pointer determines a_τ in forced-exploration rounds, and the UCB rule with a deterministic tie-breaking rule determines a_τ in adaptive-selection rounds. Hence $\boldsymbol{z}_{c,\tau}$ is $\mathcal{F}_{\tau-1}$-measurable.

Whenever $u_\tau \in \mathcal{I}_c$, the mean reward is generated by the common observable parameter $\boldsymbol{\vartheta}_c^\star$ by Eq. (3) and Definition 1:

$$y_\tau = \boldsymbol{x}_{a_\tau}^\top \boldsymbol{\vartheta}_c^\star + \epsilon_\tau.$$

Multiplying both sides by $\boldsymbol{z}_{c,\tau}$ gives, for every $\tau \in [T]$,

$$\boldsymbol{z}_{c,\tau} y_\tau = \boldsymbol{z}_{c,\tau} \boldsymbol{z}_{c,\tau}^\top \boldsymbol{\vartheta}_c^\star + \boldsymbol{z}_{c,\tau} \epsilon_\tau,$$

where both sides equal $\mathbf{0}_d$ when $u_\tau \notin \mathcal{I}_c$ because $\boldsymbol{z}_{c,\tau} = \mathbf{0}_d$. Since $\boldsymbol{z}_{c,\tau} y_\tau = \mathbf{1}\{u_\tau \in \mathcal{I}_c\} y_\tau \boldsymbol{x}_{a_\tau}$ and $\boldsymbol{z}_{c,\tau} \boldsymbol{z}_{c,\tau}^\top = \mathbf{1}\{u_\tau \in \mathcal{I}_c\} \boldsymbol{x}_{a_\tau} \boldsymbol{x}_{a_\tau}^\top$, summing over $\tau \in [t]$, for $t \le T$, and using the definitions of $\boldsymbol{b}_{\mathcal{I}_c,t}$ and $\boldsymbol{V}_{\mathcal{I}_c,t}$ yields

$$\boldsymbol{b}_{\mathcal{I}_c,t} = \left(\boldsymbol{V}_{\mathcal{I}_c,t} - \lambda \boldsymbol{I}_d \right) \boldsymbol{\vartheta}_c^\star + \sum_{\tau=1}^{t} \boldsymbol{z}_{c,\tau} \epsilon_\tau.$$

Since $\|\boldsymbol{\vartheta}_c^\star\| \le 1$ (see Section 3.2), we apply the self-normalized concentration inequality to the predictable sequence $\{\boldsymbol{z}_{c,\tau}\}_{\tau=1}^{T}$ with failure probability δ/N. Then, with probability at least $1-\delta/N$, simultaneously for all $0 \le t \le T$,

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{I}_c,t} - \boldsymbol{\vartheta}_c^\star\right\|_{\boldsymbol{V}_{\mathcal{I}_c,t}} \le \sqrt{2\log\frac{N}{\delta} + \log\frac{\det(\boldsymbol{V}_{\mathcal{I}_c,t})}{\det(\lambda \boldsymbol{I}_d)}} + \sqrt{\lambda}.$$

Since $\boldsymbol{V}_{\mathcal{I}_c,t}$ contains $T_c(t)$ observations of norm at most one,

$$\operatorname{tr}\left(\boldsymbol{V}_{\mathcal{I}_c,t} - \lambda \boldsymbol{I}_d\right) \le T_c(t) \le T.$$

The determinant-trace inequality, $T_c(t) \le T$, and the definition of β therefore imply

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{I}_c,t} - \boldsymbol{\vartheta}_c^\star\right\|_{\boldsymbol{V}_{\mathcal{I}_c,t}} \le \sqrt{2\log\frac{N}{\delta} + d\log\left(1 + \frac{T_c(t)}{\lambda d}\right)} + \sqrt{\lambda} \le \beta.$$

There are $C \le C_{\mathrm{lat}} \ll N$ observable clusters (see Section 3.2 and Lemma 3), so a union bound over the observable clusters gives failure probability at most $C\delta/N \le \delta$, establishing $\mathbb{P}(\mathcal{E}_{\mathrm{cl}}) \ge 1-\delta$. Because $\mathcal{E}_{\mathrm{cl}}$ is defined through the true observable clusters rather than through the neighborhoods used by the algorithm, the observations collected before recovery also satisfy the regression identity above and remain valid after recovery. □

Lemma 11 (Post-recovery regret using all cluster observations)**.** *Let $\mathcal{T}_{\mathrm{post}}$ be the set of adaptive-selection rounds by T in phases $s \ge s_\pi$, and let $\beta \coloneqq \sqrt{2\log(N/\delta) + d\log(1+T/(\lambda d))} + \sqrt{\lambda}$. On $\mathcal{E}_{\mathsf{e}} \cap \mathcal{E}_{\mathrm{cov}} \cap \mathcal{E}_{\mathrm{cl}}$,*

$$\mathrm{Reg}_{\mathrm{post}}(T) \coloneqq \sum_{t \in \mathcal{T}_{\mathrm{post}}} r_t \le 2\beta\sqrt{2dCT\log\left(1 + \frac{T}{\lambda d}\right)}.$$

Proof of Lemma 11. For each round t, let $c_t \coloneqq m(u_t)$. If $t \in \mathcal{T}_{\mathrm{post}}$ occurs in phase $s \ge s_\pi$, then Theorem 1 gives

$$\mathcal{N}_s(u_t) = \mathcal{I}_{c_t}.$$

The main statistics of each user contain the observations of that user from all rounds, both forced-exploration and adaptive-selection, and the pooled design $\lambda \boldsymbol{I}_d + \sum_{j \in \mathcal{N}_s(u_t)} (\boldsymbol{V}_{j,t-1} - \lambda \boldsymbol{I}_d)$ retains a single ridge term. Consequently,

$$\begin{aligned}
\boldsymbol{V}_{\mathcal{I}_{c_t},t-1} &= \lambda \boldsymbol{I}_d + \sum_{j \in \mathcal{N}_s(u_t)} (\boldsymbol{V}_{j,t-1} - \lambda \boldsymbol{I}_d), \\
\boldsymbol{b}_{\mathcal{I}_{c_t},t-1} &= \sum_{j \in \mathcal{N}_s(u_t)} \boldsymbol{b}_{j,t-1}, \\
\widehat{\boldsymbol{\theta}}_{\mathcal{I}_{c_t},t-1} &= \widehat{\boldsymbol{\theta}}_{\mathcal{N}_s(u_t),t-1}.
\end{aligned}$$

These statistics include every earlier observation of users in $\mathcal{I}_{c_t}$, including observations collected before recovery.

By Lemma 10, on $\mathcal{E}_{\mathrm{cl}}$,

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{I}_{c_t},t-1} - \boldsymbol{\vartheta}_{c_t}^\star\right\|_{\boldsymbol{V}_{\mathcal{I}_{c_t},t-1}} \le \beta.$$

For any arm k, the Cauchy-Schwarz inequality gives

$$\begin{aligned}
\left|\boldsymbol{x}_k^\top(\widehat{\boldsymbol{\theta}}_{\mathcal{I}_{c_t},t-1} - \boldsymbol{\vartheta}_{c_t}^\star)\right| &\le \|\boldsymbol{x}_k\|_{\boldsymbol{V}_{\mathcal{I}_{c_t},t-1}^{-1}} \left\|\widehat{\boldsymbol{\theta}}_{\mathcal{I}_{c_t},t-1} - \boldsymbol{\vartheta}_{c_t}^\star\right\|_{\boldsymbol{V}_{\mathcal{I}_{c_t},t-1}} \\
&\le \beta \|\boldsymbol{x}_k\|_{\boldsymbol{V}_{\mathcal{I}_{c_t},t-1}^{-1}}.
\end{aligned}$$

Recall that $a_t^\star$ is an optimal arm for cluster c_t. Applying this confidence inequality to $a_t^\star$, then using the fact that a_t maximizes the UCB score, and finally applying the lower confidence bound to a_t,

yields

$$\begin{aligned}\boldsymbol{x}_{a_t^\star}^\top \boldsymbol{\vartheta}_{c_t}^\star &\le \boldsymbol{x}_{a_t^\star}^\top \widehat{\boldsymbol{\theta}}_{\mathcal{I}_{c_t},t-1} + \beta \big\|\boldsymbol{x}_{a_t^\star}\big\|_{\boldsymbol{V}_{\mathcal{I}_{c_t},t-1}^{-1}} \\ &\le \boldsymbol{x}_{a_t}^\top \widehat{\boldsymbol{\theta}}_{\mathcal{I}_{c_t},t-1} + \beta \|\boldsymbol{x}_{a_t}\|_{\boldsymbol{V}_{\mathcal{I}_{c_t},t-1}^{-1}} \\ &\le \boldsymbol{x}_{a_t}^\top \boldsymbol{\vartheta}_{c_t}^\star + 2\beta \|\boldsymbol{x}_{a_t}\|_{\boldsymbol{V}_{\mathcal{I}_{c_t},t-1}^{-1}}.\end{aligned}$$

Subtracting the chosen arm's true mean gives

$$r_t \le 2\beta \|\boldsymbol{x}_{a_t}\|_{\boldsymbol{V}_{\mathcal{I}_{c_t},t-1}^{-1}}.$$

It remains to sum these confidence widths. For every round t, $\boldsymbol{V}_{\mathcal{I}_{c_t},t-1} \succeq \lambda \boldsymbol{I}_d$ and $\|\boldsymbol{x}_{a_t}\| \le 1$, so

$$0 \le \|\boldsymbol{x}_{a_t}\|^2_{\boldsymbol{V}_{\mathcal{I}_{c_t},t-1}^{-1}} \le \frac{1}{\lambda} \le 1,$$

and the elliptical potential lemma (Abbasi-yadkori et al., 2011) applies without truncation. For each observable cluster $c \in [C]$, let $\mathcal{T}_c \coloneqq \{t \in [T] : u_t \in \mathcal{I}_c\}$ and apply the lemma to the arms $(\boldsymbol{x}_{a_t})_{t\in\mathcal{T}_c}$. Since the set of post-recovery rounds of cluster c is a subset of $\mathcal{T}_c$ and all squared widths are nonnegative,

$$\begin{aligned}\sum_{t\in\mathcal{T}_{\mathrm{post}},c_t=c} \|\boldsymbol{x}_{a_t}\|^2_{\boldsymbol{V}_{\mathcal{I}_c,t-1}^{-1}} &\le 2\log\frac{\det(\boldsymbol{V}_{\mathcal{I}_c,T})}{\det(\lambda \boldsymbol{I}_d)} \\ &\le 2d\log\left(1+\frac{T_c(T)}{\lambda d}\right),\end{aligned}$$

where the last inequality follows from the determinant-trace inequality and the definition of $T_c(T)$ in Lemma 10. Summing over the C observable clusters and using $T_c(T) \le T$ gives

$$\sum_{t\in\mathcal{T}_{\mathrm{post}}} \|\boldsymbol{x}_{a_t}\|^2_{\boldsymbol{V}_{\mathcal{I}_{c_t},t-1}^{-1}} \le 2dC\log\left(1+\frac{T}{\lambda d}\right).$$

Finally, the Cauchy-Schwarz inequality and $|\mathcal{T}_{\mathrm{post}}| \le T$ imply

$$\begin{aligned}\mathrm{Reg}_{\mathrm{post}}(T) &\le 2\beta \sum_{t\in\mathcal{T}_{\mathrm{post}}} \|\boldsymbol{x}_{a_t}\|_{\boldsymbol{V}_{\mathcal{I}_{c_t},t-1}^{-1}} \\ &\le 2\beta\sqrt{|\mathcal{T}_{\mathrm{post}}| \sum_{t\in\mathcal{T}_{\mathrm{post}}} \|\boldsymbol{x}_{a_t}\|^2_{\boldsymbol{V}_{\mathcal{I}_{c_t},t-1}^{-1}}} \\ &\le 2\beta\sqrt{2dCT\log\left(1+\frac{T}{\lambda d}\right)}.\end{aligned}$$

□

Proof of Theorem 2. If $B = 0$, then $W = \{\mathbf{0}_d\}$ and every arm has zero mean reward for every user, so $\mathrm{Reg}(T) = 0$. Assume henceforth that $B \ge 1$. Define

$$\mathcal{E} \coloneqq \mathcal{E}_{\mathsf{e}} \cap \mathcal{E}_{\mathrm{cov}} \cap \mathcal{E}_{\mathrm{cl}}.$$

By Lemmas 6, 8 and 10, these three events fail with probabilities at most δ, $J_T\delta$, and δ, respectively. A union bound, without any independence requirement, gives

$$\mathbb{P}(\mathcal{E}) \ge 1 - (J_T + 2)\delta.$$

Every executed round belongs to exactly one of three categories: a forced-exploration round, an adaptive-selection round in a phase $s < s_\pi$, or an adaptive-selection round in a phase $s \ge s_\pi$. Therefore,

$$\mathrm{Reg}(T) = \mathrm{Reg}_{\mathrm{explore}}(T) + \mathrm{Reg}_{\mathrm{pre}}(T) + \mathrm{Reg}_{\mathrm{post}}(T).$$

By Lemma 9,

$$\mathrm{Reg}_{\mathrm{explore}}(T) \le 8C_0^p T^{1-p}, \qquad \mathrm{Reg}_{\mathrm{pre}}(T) \le 2D_\pi,$$

and by Lemma 11, on $\mathcal{E}$,

$$\mathrm{Reg}_{\mathrm{post}}(T) \le 2\beta\sqrt{2dCT\log\left(1+\frac{T}{\lambda d}\right)}.$$

Adding the three contributions yields

$$\mathrm{Reg}(T) \le 2\beta\sqrt{2dCT\log\left(1+\frac{T}{\lambda d}\right)} + 8C_0^p T^{1-p} + 2D_\pi.$$

If $C = 1$, then $s_\pi = 0$, so $D_\pi = 0$ by Lemma 9. If $C \ge 2$, substituting the upper bound on D_π in Eq. (11) gives

$$\mathrm{Reg}(T) \le 2\beta\sqrt{2dCT\log\left(1+\frac{T}{\lambda d}\right)} + 8C_0^p T^{1-p} + 8C_0\left(\frac{4}{N\pi_{\min}\Delta_{\mathcal{B}}^2}\right)^{\frac{1}{1-p}}. \tag{13}$$

It remains to express Eq. (13) in the order notation of the theorem, treating λ and p as constants. Since $\beta = \sqrt{2\log(N/\delta) + d\log(1+T/(\lambda d))} + \sqrt{\lambda} = \widetilde{\mathcal{O}}(\sqrt{d})$, the first term is $\widetilde{\mathcal{O}}(d\sqrt{CT})$. Moreover,

$$n_0^\star = B\left\lceil \frac{(2+2\sqrt{2})^2\beta^2}{B}\right\rceil \le (2+2\sqrt{2})^2\beta^2 + B = \widetilde{\mathcal{O}}(d)$$

since $B \le d$, thus

$$C_0 = \left\lceil 16N\log\frac{N}{\delta} + 4Nn_0^\star\right\rceil = \widetilde{\mathcal{O}}(Nd).$$

Hence the second term is $\widetilde{\mathcal{O}}((Nd)^p T^{1-p})$, and the third term is

$$8\cdot 4^{\frac{1}{1-p}} C_0 (N\pi_{\min})^{-\frac{1}{1-p}} \Delta_{\mathcal{B}}^{-\frac{2}{1-p}} = \widetilde{\mathcal{O}}\left(Nd(N\pi_{\min})^{-\frac{1}{1-p}}\Delta_{\mathcal{B}}^{-\frac{2}{1-p}}\right),$$

which is omitted when $C = 1$.

The concentration events above are constructed unconditionally and only then intersected; in particular, the analysis never conditions the noise distribution on successful recovery. For a desired overall failure probability $\eta \in (0,1)$, setting

$$\delta = \frac{\eta}{J_T + 2}$$

in the algorithmic calibration ensures total failure probability at most η without using any arrival probability. □

D Extension to Time-varying Contexts

We extend observable clustering to IID time-varying candidate sets without prior knowledge of the population feature span. For the analysis, we slightly modify Algorithm 1: the learner incrementally expands a basis to cover the span of observed action features and dynamically pads the regression statistics to preserve the past information. Furthermore, uniform exploration replaces spanning-block exploration, while the remaining mechanisms of Algorithm 1 are retained.

D.1 Problem Formulation

At round t, the learner observes a user u_t and an arm set $\mathcal{A}_t = \{\boldsymbol{x}_{t,k} : k \in [K]\}$ before selecting an index $a_t \in [K]$. The feature vectors in $\mathcal{A}_t$ are drawn IID from an unknown distribution $\mathcal{P}$, independently of u_t and of the past, and satisfy $\|\boldsymbol{x}_{t,k}\| \le 1$ for all $t \in [T]$ and $k \in [K]$. Define a filtration $\mathcal{F}_{t-1} := \sigma\left((u_\tau, \boldsymbol{x}_{\tau,a_\tau}, y_\tau)_{\tau<t}, u_t, \boldsymbol{x}_{t,a_t}\right)$, and let ϵ_t be conditionally 1-sub-Gaussian given $\mathcal{F}_{t-1}$. The cumulative regret is $\mathrm{Reg}(T) := \sum_{t=1}^T (\boldsymbol{x}_{t,a_t^\star}^\top \boldsymbol{\theta}_{u_t}^\star - \boldsymbol{x}_{t,a_t}^\top \boldsymbol{\theta}_{u_t}^\star)$, where $a_t^\star \in \mathrm{argmax}_{k\in[K]} \boldsymbol{x}_{t,k}^\top \boldsymbol{\theta}_{u_t}^\star$.

All other elements of the environment, including reward generation, the user-arrival model, and the bound $\|\boldsymbol{\theta}_i^\star\| \le 1$ in Assumption 2, follow the setup in Section 3.2.

Let $\boldsymbol{\Sigma}_{\mathcal{P}} \coloneqq \mathbb{E}_{\boldsymbol{u}\sim\mathcal{P}}[\boldsymbol{u}\boldsymbol{u}^\top]$ denote the population second-moment matrix of the feature-generating distribution $\mathcal{P}$, and let $W_{\mathcal{P}} \coloneqq \mathrm{range}(\boldsymbol{\Sigma}_{\mathcal{P}})$ denote the corresponding population feature span. Note that a single arm set $\mathcal{A}_t$ need not span $W_{\mathcal{P}}$. Let $\Pi_{W_{\mathcal{P}}}$ denote the orthogonal projection onto $W_{\mathcal{P}}$ and define the population observable partition by

$$\begin{aligned} i \sim_{\mathcal{P}} j &\iff \Pi_{W_{\mathcal{P}}}\boldsymbol{\theta}_i^\star = \Pi_{W_{\mathcal{P}}}\boldsymbol{\theta}_j^\star \\ &\iff \boldsymbol{u}^\top\boldsymbol{\theta}_i^\star = \boldsymbol{u}^\top\boldsymbol{\theta}_j^\star, \end{aligned} \tag{14}$$

for $\mathcal{P}$-almost every $\boldsymbol{u} \in \mathbb{R}^d$. Indeed, for $\boldsymbol{v} \coloneqq \boldsymbol{\theta}_i^\star - \boldsymbol{\theta}_j^\star$, it holds that

$$\boldsymbol{v}^\top\boldsymbol{\Sigma}_{\mathcal{P}}\boldsymbol{v} = \mathbb{E}_{\boldsymbol{u}\sim\mathcal{P}}\left[(\boldsymbol{u}^\top\boldsymbol{v})^2\right], \tag{15}$$

$$\boldsymbol{\Sigma}_{\mathcal{P}}\boldsymbol{v} = \mathbb{E}_{\boldsymbol{u}\sim\mathcal{P}}\left[\boldsymbol{u}(\boldsymbol{u}^\top\boldsymbol{v})\right]. \tag{16}$$

Since $(\boldsymbol{u}^\top\boldsymbol{v})^2 \ge 0$, Eq. (15) shows that $\boldsymbol{u}^\top\boldsymbol{v} = 0$ holds $\mathcal{P}$-almost surely if and only if $\boldsymbol{v}^\top\boldsymbol{\Sigma}_{\mathcal{P}}\boldsymbol{v} = 0$. If $\boldsymbol{u}^\top\boldsymbol{v} = 0$ holds $\mathcal{P}$-almost surely, Eq. (16) gives $\boldsymbol{\Sigma}_{\mathcal{P}}\boldsymbol{v} = \boldsymbol{0}_d$, and conversely, $\boldsymbol{\Sigma}_{\mathcal{P}}\boldsymbol{v} = \boldsymbol{0}_d$ implies $\boldsymbol{v}^\top\boldsymbol{\Sigma}_{\mathcal{P}}\boldsymbol{v} = 0$. Hence, the conditions $\boldsymbol{u}^\top\boldsymbol{v} = 0$ $\mathcal{P}$-almost surely, $\boldsymbol{v}^\top\boldsymbol{\Sigma}_{\mathcal{P}}\boldsymbol{v} = 0$, and $\boldsymbol{\Sigma}_{\mathcal{P}}\boldsymbol{v} = \boldsymbol{0}_d$ are equivalent. Since $\boldsymbol{\Sigma}_{\mathcal{P}}$ is symmetric, $\ker(\boldsymbol{\Sigma}_{\mathcal{P}}) = \mathrm{range}(\boldsymbol{\Sigma}_{\mathcal{P}})^\perp = W_{\mathcal{P}}^\perp$. The condition $\boldsymbol{\Sigma}_{\mathcal{P}}\boldsymbol{v} = \boldsymbol{0}_d$ is therefore equivalent to $\Pi_{W_{\mathcal{P}}}\boldsymbol{\theta}_i^\star = \Pi_{W_{\mathcal{P}}}\boldsymbol{\theta}_j^\star$, proving Eq. (14). We denote the number of population observable clusters by

$$C_{\mathcal{P}} \coloneqq |\{\Pi_{W_{\mathcal{P}}}\boldsymbol{\theta}_i^\star : i \in [N]\}|,$$

where $C_{\mathcal{P}} \le C_{\mathrm{lat}}$ follows from the argument of Lemma 3 with W replaced by $W_{\mathcal{P}}$. Throughout this section, let $\{\mathcal{I}_c\}_{c=1}^{C_{\mathcal{P}}}$ denote these clusters and $\boldsymbol{\vartheta}_c^\star$ their common projected parameters. Define the cluster gap as

$$\Delta_{\mathcal{P}} \coloneqq \min_{i \not\sim_{\mathcal{P}} j} \left\|\boldsymbol{\theta}_i^\star - \boldsymbol{\theta}_j^\star\right\|_{\boldsymbol{\Sigma}_{\mathcal{P}}}.$$

By Eq. (14), every pair of users with $i \not\sim_{\mathcal{P}} j$ satisfies $\|\boldsymbol{\theta}_i^\star - \boldsymbol{\theta}_j^\star\|_{\boldsymbol{\Sigma}_{\mathcal{P}}} > 0$, and the minimum is taken over finitely many such pairs, so $\Delta_{\mathcal{P}} > 0$. When $C_{\mathcal{P}} = 1$, no gap is needed and $\Delta_{\mathcal{P}}$ is left undefined. The learner knows neither the distribution and its span nor the partition, its cardinality, the gap, or the arrival probabilities.

D.2 Algorithm

Unlike the main setting, in which the fixed arm set $\mathcal{A}$ is given, the learner does not know the population feature span $W_{\mathcal{P}}$ in advance. We therefore employ an online span-expansion procedure that tracks the span of all feature vectors observed up to the current round. Define $\widehat{W}_0 \coloneqq \{\boldsymbol{0}_d\}$ and, for $t \ge 1$,

$$\widehat{W}_t \coloneqq \mathrm{span}\left\{\boldsymbol{x}_{\tau,k} : 1 \le \tau \le t, k \in [K]\right\},$$

and let $\widehat{d}_t \coloneqq \dim(\widehat{W}_t)$ denote the dimension of the observed feature span at round t, so that $\widehat{d}_0 = 0$. The learner represents $\widehat{W}_t$ by a matrix $\boldsymbol{U}_t \in \mathbb{R}^{d\times\widehat{d}_t}$ whose columns form an orthonormal basis of $\widehat{W}_t$, starting from the empty matrix $\boldsymbol{U}_0 \in \mathbb{R}^{d\times 0}$. Since the whole arm set $\mathcal{A}_t$ is revealed before selection, the learner obtains $\boldsymbol{U}_t$ before selecting a_t by Gram-Schmidt orthogonalization of the features in $\mathcal{A}_t$. Specifically, starting from $\boldsymbol{U}_t^{(0)} \coloneqq \boldsymbol{U}_{t-1}$, for $k \in [K]$, let

$$\boldsymbol{v}_{t,k} \coloneqq \left(\boldsymbol{I}_d - \boldsymbol{U}_t^{(k-1)}\boldsymbol{U}_t^{(k-1)\top}\right)\boldsymbol{x}_{t,k},$$

and construct each $\boldsymbol{U}_t^{(k)}$ following

$$\boldsymbol{U}_t^{(k)} \coloneqq \begin{cases} \left[\boldsymbol{U}_t^{(k-1)} \;\; \boldsymbol{v}_{t,k}/\|\boldsymbol{v}_{t,k}\|\right] & \text{if } \boldsymbol{v}_{t,k} \ne \boldsymbol{0}_d, \\ \boldsymbol{U}_t^{(k-1)} & \text{otherwise,} \end{cases}$$

and set $\boldsymbol{U}_t \coloneqq \boldsymbol{U}_t^{(K)}$. Since each $\boldsymbol{v}_{t,k}$ is orthogonal to every column of $\boldsymbol{U}_t^{(k-1)}$, the columns of $\boldsymbol{U}_t$ remain orthonormal, and all columns of $\boldsymbol{U}_{t-1}$ remain unchanged. After processing all K features, the columns of $\boldsymbol{U}_t$ span $\widehat{W}_t$.

Before presenting the algorithm for this setting, we establish that the observed spans are a nested sequence of subspaces within $W_{\mathcal{P}}$, ensuring that their dimensions do not exceed $\dim(W_{\mathcal{P}})$.

Lemma 12 (Dimension of the observed feature span)**.** *With probability one, simultaneously for all* $t \in [T]$*,*

$$\widehat{W}_{t-1} \subseteq \widehat{W}_t \subseteq W_{\mathcal{P}}, \quad \textit{and} \quad 0 \le \widehat{d}_{t-1} \le \widehat{d}_t \le \dim(W_{\mathcal{P}}) \le d.$$

Consequently, at most $\dim(W_{\mathcal{P}})$ *basis columns are appended during the* T *rounds.*

Proof of Lemma 12. Note that $\Pi_{W_{\mathcal{P}}}$ is an orthogonal projection, thus symmetric and idempotent. Consequently, it holds that

$$(\boldsymbol{I}_d - \Pi_{W_{\mathcal{P}}})^\top (\boldsymbol{I}_d - \Pi_{W_{\mathcal{P}}}) = \boldsymbol{I}_d - \Pi_{W_{\mathcal{P}}}.$$

Expanding the squared norm and using that $\boldsymbol{u}^\top \boldsymbol{M}\boldsymbol{u} = \mathrm{tr}(\boldsymbol{M}\boldsymbol{u}\boldsymbol{u}^\top)$ for any $\boldsymbol{M} \in \mathbb{R}^{d\times d}$, we obtain

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{u}\sim\mathcal{P}}\left[\|(\boldsymbol{I}_d - \Pi_{W_{\mathcal{P}}})\boldsymbol{u}\|^2\right] &= \mathbb{E}_{\boldsymbol{u}\sim\mathcal{P}}\left[\boldsymbol{u}^\top(\boldsymbol{I}_d - \Pi_{W_{\mathcal{P}}})\boldsymbol{u}\right] \\
&= \mathbb{E}_{\boldsymbol{u}\sim\mathcal{P}}\left[\mathrm{tr}((\boldsymbol{I}_d - \Pi_{W_{\mathcal{P}}})\boldsymbol{u}\boldsymbol{u}^\top)\right] \\
&= \mathrm{tr}\left((\boldsymbol{I}_d - \Pi_{W_{\mathcal{P}}})\mathbb{E}_{\boldsymbol{u}\sim\mathcal{P}}[\boldsymbol{u}\boldsymbol{u}^\top]\right) \qquad (17)\\
&= \mathrm{tr}\left((\boldsymbol{I}_d - \Pi_{W_{\mathcal{P}}})\boldsymbol{\Sigma}_{\mathcal{P}}\right) \\
&= 0, \qquad (18)
\end{aligned}$$

where Eq. (17) interchanges the expectation and the trace by linearity, which is valid because the arm features are bounded. Eq. (18) holds because every column of $\boldsymbol{\Sigma}_{\mathcal{P}}$ lies in $W_{\mathcal{P}} = \mathrm{range}(\boldsymbol{\Sigma}_{\mathcal{P}})$, so that $\Pi_{W_{\mathcal{P}}}\boldsymbol{\Sigma}_{\mathcal{P}} = \boldsymbol{\Sigma}_{\mathcal{P}}$.

Because the squared norm is nonnegative, Eq. (18) implies that $(\boldsymbol{I}_d - \Pi_{W_{\mathcal{P}}})\boldsymbol{u} = \boldsymbol{0}_d$ holds $\mathcal{P}$-almost surely, that is, $\boldsymbol{u} \in W_{\mathcal{P}}$ for $\mathcal{P}$-almost every $\boldsymbol{u}$. Moreover, since $\boldsymbol{x}_{t,k} \sim \mathcal{P}$ for every $t \in [T]$ and $k \in [K]$, a union bound gives

$$\begin{aligned}
\mathbb{P}\left(\exists\, t \in [T],\ k \in [K] : \boldsymbol{x}_{t,k} \notin W_{\mathcal{P}}\right) &\le \sum_{t=1}^{T}\sum_{k=1}^{K} \mathbb{P}\left(\boldsymbol{x}_{t,k} \notin W_{\mathcal{P}}\right) \\
&= 0.
\end{aligned}$$

Hence, with probability one, all arm features belong to $W_{\mathcal{P}}$ simultaneously.

On this event, $\widehat{W}_t \subseteq W_{\mathcal{P}}$ for every $t \in [T]$, because $\widehat{W}_t$ is spanned by vectors in $W_{\mathcal{P}}$. Moreover, the features defining $\widehat{W}_{t-1}$ are included among those defining $\widehat{W}_t$, so $\widehat{W}_{t-1} \subseteq \widehat{W}_t$. Since dimension is monotone under subspace inclusion, the preceding inclusions imply, for all $t \in [T]$,

$$0 \le \widehat{d}_{t-1} \le \widehat{d}_t \le \dim(W_{\mathcal{P}}) \le d.$$

Finally, $\boldsymbol{U}_t \in \mathbb{R}^{d\times \widehat{d}_t}$ is obtained by appending columns to $\boldsymbol{U}_{t-1} \in \mathbb{R}^{d\times \widehat{d}_{t-1}}$, so exactly $\widehat{d}_t - \widehat{d}_{t-1}$ columns are appended at round t. Since $\widehat{d}_0 = 0$, the total number appended by time T is

$$\sum_{t=1}^{T}(\widehat{d}_t - \widehat{d}_{t-1}) = \widehat{d}_T - \widehat{d}_0 = \widehat{d}_T \le \dim(W_{\mathcal{P}}) \le d,$$

which completes the proof. □

With Lemma 12 established, we present Algorithm 4. We set $\lambda = 1$ and maintain the regression statistics in the coordinates of the evolving basis $\boldsymbol{U}_t$. Specifically, for each user i, we write $(\bar{\boldsymbol{V}}_i, \bar{\boldsymbol{b}}_i)$ for the main statistics, $(\bar{\boldsymbol{V}}_{i,\mathrm{e}}, \bar{\boldsymbol{b}}_{i,\mathrm{e}})$ for the exploration statistics, and $n_{i,\mathrm{e}}$ for the exploration count, suppressing the round index. Whenever $q_t := \widehat{d}_t - \widehat{d}_{t-1} > 0$ new columns are appended at round t, both statistics pairs of every user are padded as

$$\bar{\boldsymbol{V}} \leftarrow \begin{pmatrix} \bar{\boldsymbol{V}} & \boldsymbol{0}_{\widehat{d}_{t-1}\times q_t} \\ \boldsymbol{0}_{q_t\times \widehat{d}_{t-1}} & \boldsymbol{I}_{q_t} \end{pmatrix}, \qquad \bar{\boldsymbol{b}} \leftarrow \begin{pmatrix} \bar{\boldsymbol{b}} \\ \boldsymbol{0}_{q_t} \end{pmatrix}. \qquad (19)$$

The new columns are orthogonal to $\widehat{W}_{t-1}$, and every previously selected feature lies in $\widehat{W}_{t-1}$, so every past observation has zero coordinates along the new columns. Hence past observations contribute nothing to the new rows and columns of $\bar{\boldsymbol{V}}$ or to the new entries of $\bar{\boldsymbol{b}}$, the new diagonal block $\boldsymbol{I}_{q_t}$ is

Algorithm 4 TRIBE with IID time-varying action sets and online span expansion

Input: Number of arms K, regularization parameter $\lambda = 1$, confidence parameter $\delta \in (0, 1)$, phase parameter $p \in (1/2, 1)$, UCB coefficient β, phase-scale constant C_0
Initialize an empty basis $\boldsymbol{U}_0 \in \mathbb{R}^{d\times 0}$ and set $\widehat{d}_0 \leftarrow 0$
Initialize the complete graph $\mathcal{G}_{-1} = ([N], E(\mathcal{G}_{-1}))$
For all $i \in [N]$,
    set main statistics $\bar{\boldsymbol{V}}_{i,0} \leftarrow \lambda \boldsymbol{I}_0$, $\bar{\boldsymbol{b}}_{i,0} \leftarrow \boldsymbol{0}_0$
    set exploration statistics $\bar{\boldsymbol{V}}_{i,\mathrm{e},0} \leftarrow \lambda \boldsymbol{I}_0$, $\bar{\boldsymbol{b}}_{i,\mathrm{e},0} \leftarrow \boldsymbol{0}_0$, $n_{i,\mathrm{e},0} \leftarrow 0$
$t \leftarrow 0$
for $s = 0, 1, 2, \ldots$ **do**
    $M_s \leftarrow C_0 2^s, \quad L_s \leftarrow \lceil 2^{ps/(1-p)} M_s \rceil$
    for $\ell = 1, \ldots, M_s + L_s$ **(terminate the algorithm if** $t = T$**) do**
        $t \leftarrow t + 1$
        Observe current user u_t and arm set $\mathcal{A}_t$
        PREPARE-ROUND(t)
        if $\ell \le M_s$ **then** ▷ *Forced exploration*
            Draw $a_t \sim \mathrm{Unif}([K])$
            $y_t \leftarrow$ UPDATE-IID-EXPLORATION(t, u_t, a_t)
        else ▷ *Adaptive selection*
            if $\ell = M_s + 1$ **then** ▷ *Update graph once*
                $(\mathcal{G}_s, \{\mathcal{N}_s(i)\}_{i=1}^N) \leftarrow$ HM-EDGETEST$(\mathcal{G}_{s-1}, \{\bar{\boldsymbol{V}}_{i,\mathrm{e},t-1}, \bar{\boldsymbol{b}}_{i,\mathrm{e},t-1}, n_{i,\mathrm{e},t-1}\}_{i=1}^N, \lambda, \delta/4)$
            end if
            $\bar{\boldsymbol{V}}_{\mathcal{N}_s(u_t),t-1} \leftarrow \lambda \boldsymbol{I}_{\widehat{d}_t} + \sum_{j\in\mathcal{N}_s(u_t)} \left(\bar{\boldsymbol{V}}_{j,t-1} - \lambda \boldsymbol{I}_{\widehat{d}_t}\right)$
            $\bar{\boldsymbol{b}}_{\mathcal{N}_s(u_t),t-1} \leftarrow \sum_{j\in\mathcal{N}_s(u_t)} \bar{\boldsymbol{b}}_{j,t-1}$
            $\widehat{\bar{\boldsymbol{\theta}}}_{\mathcal{N}_s(u_t),t-1} \leftarrow \bar{\boldsymbol{V}}^{-1}_{\mathcal{N}_s(u_t),t-1} \bar{\boldsymbol{b}}_{\mathcal{N}_s(u_t),t-1}$
            Choose $a_t \in \operatorname*{argmax}_{k\in[K]} \left\{ \boldsymbol{z}_{t,k}^\top \widehat{\bar{\boldsymbol{\theta}}}_{\mathcal{N}_s(u_t),t-1} + \beta \|\boldsymbol{z}_{t,k}\|_{\bar{\boldsymbol{V}}^{-1}_{\mathcal{N}_s(u_t),t-1}} \right\}$
            Play arm a_t and observe y_t
        end if
        $\bar{\boldsymbol{V}}_{u_t,t} \leftarrow \bar{\boldsymbol{V}}_{u_t,t-1} + \boldsymbol{z}_{t,a_t}\boldsymbol{z}_{t,a_t}^\top, \quad \bar{\boldsymbol{b}}_{u_t,t} \leftarrow \bar{\boldsymbol{b}}_{u_t,t-1} + y_t \boldsymbol{z}_{t,a_t}$
    end for
end for

Algorithm 5 Prepare-Round

procedure PREPARE-ROUND(t)
    $\boldsymbol{U}_t \leftarrow \boldsymbol{U}_{t-1}, \quad q_t \leftarrow 0$
    for $k = 1, \ldots, K$ **do**
        $\boldsymbol{v} \leftarrow \boldsymbol{x}_{t,k} - \boldsymbol{U}_t \boldsymbol{U}_t^\top \boldsymbol{x}_{t,k}$
        if $\|\boldsymbol{v}\| > 0$ **then**
            $\boldsymbol{U}_t \leftarrow [\boldsymbol{U}_t \ \ \boldsymbol{v}/\|\boldsymbol{v}\|], \quad q_t \leftarrow q_t + 1$
        end if
    end for
    $\widehat{d}_t \leftarrow$ number of columns of $\boldsymbol{U}_t$
    if $q_t > 0$ **then**
        For all $i \in [N]$, pad $(\bar{\boldsymbol{V}}_{i,t-1}, \bar{\boldsymbol{b}}_{i,t-1})$ and $(\bar{\boldsymbol{V}}_{i,\mathrm{e},t-1}, \bar{\boldsymbol{b}}_{i,\mathrm{e},t-1})$ according to Eq. (19)
    end if
    $\boldsymbol{z}_{t,k} \leftarrow \boldsymbol{U}_t^\top \boldsymbol{x}_{t,k}, \quad \forall k \in [K]$
end procedure

the ridge term $\lambda \boldsymbol{I}_{q_t}$ with $\lambda = 1$, and the exploration counts remain unchanged. Since Algorithm 5 is executed before the graph update, the exploration statistics passed to Algorithm 2 are already padded to $\widehat{d}_t$ coordinates; in Appendix D.3, we demonstrate that the padding leaves the outcome of the separation test unchanged.

Furthermore, the spanning-block exploration in Algorithm 1 is replaced by uniform exploration over arm indices, as in Li et al. (2025), and each exploration observation is incorporated immediately into

Algorithm 6 `Update-IID-Exploration`

function UPDATE-IID-EXPLORATION(t, u_t, a_t)
  Play arm a_t and observe y_t
  $\bar{\boldsymbol{V}}_{u_t,\mathrm{e},t} \leftarrow \bar{\boldsymbol{V}}_{u_t,\mathrm{e},t-1} + \boldsymbol{z}_{t,a_t}\boldsymbol{z}_{t,a_t}^\top$
  $\bar{\boldsymbol{b}}_{u_t,\mathrm{e},t} \leftarrow \bar{\boldsymbol{b}}_{u_t,\mathrm{e},t-1} + y_t\boldsymbol{z}_{t,a_t}$
  $n_{u_t,\mathrm{e},t} \leftarrow n_{u_t,\mathrm{e},t-1} + 1$
  return y_t
end function

the corresponding user's exploration regression statistics, without block commitment. The phase schedule of Algorithm 1 and the harmonic-mean separation test in Algorithm 2 are otherwise retained, with the confidence parameter $\delta/4$ in place of δ. Accordingly, for every user i and round t, the exploration radius computed in Algorithm 2 with $\lambda = 1$ is

$$\beta_{i,\mathrm{e},t} \coloneqq \sqrt{2\log\frac{4N}{\delta} + d\log\left(1 + \frac{n_{i,\mathrm{e},t}}{d}\right)} + 1.$$

At the graph update of phase s, executed at round $t_s + 1$ with the exploration statistics through round t_s, define the exploration estimator $\widehat{\bar{\boldsymbol{\theta}}}_{i,\mathrm{e},t_s} \coloneqq \bar{\boldsymbol{V}}_{i,\mathrm{e},t_s}^{-1}\bar{\boldsymbol{b}}_{i,\mathrm{e},t_s}$ and the harmonic-mean matrix $\bar{\boldsymbol{H}}_{ij,t_s} \coloneqq (\bar{\boldsymbol{V}}_{i,\mathrm{e},t_s}^{-1} + \bar{\boldsymbol{V}}_{j,\mathrm{e},t_s}^{-1})^{-1}$. Then, an existing edge (i, j) is deleted exactly when

$$\left\|\widehat{\bar{\boldsymbol{\theta}}}_{i,\mathrm{e},t_s} - \widehat{\bar{\boldsymbol{\theta}}}_{j,\mathrm{e},t_s}\right\|_{\bar{\boldsymbol{H}}_{ij,t_s}} > \sqrt{\beta_{i,\mathrm{e},t_s}^2 + \beta_{j,\mathrm{e},t_s}^2}. \tag{20}$$

D.3 EQUIVALENCE OF FEATURE REPRESENTATIONS

Algorithm 4 stores and updates the regression statistics in the coordinates of the evolving basis $\boldsymbol{U}_t$, whereas the analysis in Appendix D.4 works with ridge statistics formed from the original arm features in $\mathbb{R}^d$. The analysis requires the latter representation because the self-normalized concentration bound (Abbasi-yadkori et al., 2011) applies to predictable regressors of a fixed dimension and to a fixed parameter. In the reduced coordinates, the dimension $\widehat{d}_t$ grows over time, and expressing all regressors in a single basis would require the final basis $\boldsymbol{U}_T$, which depends on future arm sets; in $\mathbb{R}^d$, the dimension d and the projected parameters $\Pi_{W_\mathcal{P}}\boldsymbol{\theta}_i^\star$ remain fixed. This section shows that the two representations yield identical predicted rewards, confidence widths, and harmonic-mean separation statistics, so that Algorithm 4 selects the same arms and deletes the same edges as the variant that runs ridge regression with $\lambda = 1$ on the original arm features in $\mathbb{R}^d$.

Fix $t \in [T]$ and let $q_t \coloneqq \widehat{d}_t - \widehat{d}_{t-1}$ be the number of columns appended at round t. Since $\boldsymbol{U}_t$ is obtained by appending columns to $\boldsymbol{U}_{t-1}$, we write $\boldsymbol{U}_t = [\boldsymbol{U}_{t-1}\ \boldsymbol{Q}_t]$, where $\boldsymbol{Q}_t \in \mathbb{R}^{d\times q_t}$ collects the appended columns. Because the columns of $\boldsymbol{U}_t$ are orthonormal, $\boldsymbol{Q}_t^\top\boldsymbol{U}_{t-1} = \boldsymbol{0}_{q_t\times\widehat{d}_{t-1}}$ and $\boldsymbol{Q}_t^\top\boldsymbol{Q}_t = \boldsymbol{I}_{q_t}$; in particular, the columns of $\boldsymbol{Q}_t$ are orthogonal to $\widehat{W}_{t-1}$. When $q_t = 0$, $\boldsymbol{Q}_t$ has no columns and no padding is needed.

Every previously selected feature $\boldsymbol{x}_{\tau,a_\tau}$ with $1 \le \tau < t$ lies in $\widehat{W}_{t-1}$ and is therefore orthogonal to the columns of $\boldsymbol{Q}_t$, so $\boldsymbol{Q}_t^\top\boldsymbol{x}_{\tau,a_\tau} = \boldsymbol{0}_{q_t}$. Consequently,

$$\boldsymbol{U}_t^\top\boldsymbol{x}_{\tau,a_\tau} = \begin{pmatrix}\boldsymbol{U}_{t-1}^\top\boldsymbol{x}_{\tau,a_\tau}\\ \boldsymbol{0}_{q_t}\end{pmatrix},$$

for $1 \le \tau < t$. Thus, expanding the basis leaves the existing coordinates of every past selected feature unchanged and appends zeros in the new coordinates. Therefore, after the padding in Eq. (19), the main and exploration statistics of every user equal the statistics constructed directly from the same past observations in the coordinates of $\boldsymbol{U}_t$.

We next show that computations based on the original arm features $\boldsymbol{x}_{t,k} \in \mathbb{R}^d$ and those based on their reduced coordinates $\boldsymbol{z}_{t,k} = \boldsymbol{U}_t^\top\boldsymbol{x}_{t,k}$ yield the same predicted rewards, confidence widths, and harmonic-mean test statistics.

Fix $t \in [T]$ after basis expansion and define $\boldsymbol{P}_t \coloneqq \boldsymbol{I}_d - \boldsymbol{U}_t\boldsymbol{U}_t^\top \in \mathbb{R}^{d\times d}$, the orthogonal projection onto $\widehat{W}_t^\perp$. In particular, $\boldsymbol{P}_t$ is symmetric and idempotent, with $\boldsymbol{P}_t\boldsymbol{U}_t = \boldsymbol{0}_{d\times\widehat{d}_t}$. Consider any

collection $\mathcal{S}$ of observations $(\boldsymbol{x}, y)$ whose features lie in $\widehat{W}_t$, such as the observations collected through round t by a user or by a neighborhood, and define the statistics constructed from the original arm features by

$$\boldsymbol{V} \coloneqq \boldsymbol{I}_d + \sum_{(\boldsymbol{x},y)\in\mathcal{S}} \boldsymbol{x}\boldsymbol{x}^\top, \quad \boldsymbol{b} \coloneqq \sum_{(\boldsymbol{x},y)\in\mathcal{S}} y\boldsymbol{x}$$

and from their coordinates with respect to $\boldsymbol{U}_t$ by

$$\bar{\boldsymbol{V}} \coloneqq \boldsymbol{I}_{\widehat{d}_t} + \sum_{(\boldsymbol{x},y)\in\mathcal{S}} \boldsymbol{U}_t^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{U}_t, \quad \bar{\boldsymbol{b}} \coloneqq \sum_{(\boldsymbol{x},y)\in\mathcal{S}} y\,\boldsymbol{U}_t^\top \boldsymbol{x}.$$

Since every feature in $\mathcal{S}$ satisfies $\boldsymbol{U}_t\boldsymbol{U}_t^\top \boldsymbol{x} = \boldsymbol{x}$, we have

$$\boldsymbol{V} = \boldsymbol{U}_t \bar{\boldsymbol{V}} \boldsymbol{U}_t^\top + \boldsymbol{P}_t, \quad \boldsymbol{b} = \boldsymbol{U}_t \bar{\boldsymbol{b}}, \tag{21}$$

$$\boldsymbol{V}^{-1} = \boldsymbol{U}_t \bar{\boldsymbol{V}}^{-1} \boldsymbol{U}_t^\top + \boldsymbol{P}_t. \tag{22}$$

The term $\boldsymbol{U}_t\bar{\boldsymbol{V}}\boldsymbol{U}_t^\top$ contains the data contributions and regularization on $\widehat{W}_t$, whereas $\boldsymbol{P}_t$ accounts for regularization on $\widehat{W}_t^\perp$, where the observations contribute no information. The inverse identity, Eq. (22), follows from $\boldsymbol{U}_t^\top \boldsymbol{U}_t = \boldsymbol{I}_{\widehat{d}_t}$ and the projection properties above.

Every currently available arm feature satisfies $\boldsymbol{x}_{t,k} = \boldsymbol{U}_t \boldsymbol{z}_{t,k}$. Substituting Eq. (21) and Eq. (22) therefore gives $\boldsymbol{x}_{t,k}^\top \boldsymbol{V}^{-1}\boldsymbol{b} = \boldsymbol{z}_{t,k}^\top \bar{\boldsymbol{V}}^{-1}\bar{\boldsymbol{b}}$ and $\|\boldsymbol{x}_{t,k}\|_{\boldsymbol{V}^{-1}} = \|\boldsymbol{z}_{t,k}\|_{\bar{\boldsymbol{V}}^{-1}}$. These identities also apply to the neighborhood-pooled statistics, which are constructed from the pooled observations of the neighborhood. Hence the two representations yield the same predicted rewards, confidence widths, and UCB scores for every arm, and therefore select the same arm under the same tie-breaking rule.

At the graph update of phase s, Algorithm 5 has already been executed at round $t_s + 1$, so the exploration statistics through round t_s are expressed in the coordinates of $\boldsymbol{U}_{t_s+1}$. Since $\widehat{W}_{t_s} \subseteq \widehat{W}_{t_s+1}$, the exploration observations of each user i through round t_s constitute a collection $\mathcal{S}$ as above with $t = t_s + 1$, and Eq. (21) and Eq. (22) give

$$\boldsymbol{V}_{i,\mathsf{e},t_s}^{-1} = \boldsymbol{U}_{t_s+1} \bar{\boldsymbol{V}}_{i,\mathsf{e},t_s}^{-1} \boldsymbol{U}_{t_s+1}^\top + \boldsymbol{P}_{t_s+1}, \qquad \boldsymbol{b}_{i,\mathsf{e},t_s} = \boldsymbol{U}_{t_s+1} \bar{\boldsymbol{b}}_{i,\mathsf{e},t_s}.$$

By Eq. (22),

$$\boldsymbol{V}_{i,\mathsf{e},t_s}^{-1} + \boldsymbol{V}_{j,\mathsf{e},t_s}^{-1} = \boldsymbol{U}_{t_s+1} \left(\bar{\boldsymbol{V}}_{i,\mathsf{e},t_s}^{-1} + \bar{\boldsymbol{V}}_{j,\mathsf{e},t_s}^{-1} \right) \boldsymbol{U}_{t_s+1}^\top + 2\boldsymbol{P}_{t_s+1},$$

and the same argument as for Eq. (22) yields

$$\boldsymbol{H}_{ij,t_s} = \boldsymbol{U}_{t_s+1} \bar{\boldsymbol{H}}_{ij,t_s} \boldsymbol{U}_{t_s+1}^\top + \frac{1}{2}\boldsymbol{P}_{t_s+1}.$$

Moreover, the original-feature and reduced-coordinate estimators satisfy

$$\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s} \coloneqq \boldsymbol{V}_{i,\mathsf{e},t_s}^{-1} \boldsymbol{b}_{i,\mathsf{e},t_s} = \boldsymbol{U}_{t_s+1} \widehat{\bar{\boldsymbol{\theta}}}_{i,\mathsf{e},t_s}$$

for every $i \in [N]$. The difference between any two original-feature estimators lies in $\widehat{W}_{t_s+1}$ and is therefore mapped to zero by $\boldsymbol{P}_{t_s+1}$. Consequently,

$$\left\| \widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s} - \widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s} \right\|_{\boldsymbol{H}_{ij,t_s}} = \left\| \widehat{\bar{\boldsymbol{\theta}}}_{i,\mathsf{e},t_s} - \widehat{\bar{\boldsymbol{\theta}}}_{j,\mathsf{e},t_s} \right\|_{\bar{\boldsymbol{H}}_{ij,t_s}}.$$

Thus the harmonic-mean test statistics are identical in the two representations. The thresholds depend only on the exploration counts, which do not depend on the representation, so both representations delete the same edges.

D.4 Regret Analysis

Using the equivalence established in Appendix D.3, we analyze the regression statistics formed from the original arm features with $\lambda = 1$. In the main setting, block commitment makes each exploration design matrix a deterministic function of the exploration count, $\boldsymbol{V}_{i,\mathsf{e},t} = \lambda \boldsymbol{I}_d + n_{i,\mathsf{e},t}\boldsymbol{\Sigma}_{\mathcal{B}}$. Under uniform exploration, the exploration design matrices also depend on the random arm features and arm draws, so this identity is no longer available. The following lemma instead gives a lower bound, in the Loewner order, on the identity matrix plus a sum of independent random positive-semidefinite matrices.

Lemma 13 (Regularized matrix lower bound)**.** *Let $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_m$ be independent random symmetric $d \times d$ matrices with $\mathbf{0}_{d\times d} \preceq \boldsymbol{Y}_j \preceq \boldsymbol{I}_d$ almost surely, and set $\boldsymbol{G} \coloneqq \sum_{j=1}^m \mathbb{E}[\boldsymbol{Y}_j]$. For any integer $\ell \geq 1$, with probability at least $1 - d\exp(-\ell)$,*

$$\boldsymbol{I}_d + \sum_{j=1}^m \boldsymbol{Y}_j \succeq \frac{1}{2}\boldsymbol{I}_d + \frac{\boldsymbol{G}}{16\ell}.$$

Proof of Lemma 13. Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the minimum and maximum eigenvalues of a symmetric matrix. Note that $\boldsymbol{G} = \sum_{j=1}^m \mathbb{E}[\boldsymbol{Y}_j]$ is a deterministic positive-semidefinite matrix. Let $\boldsymbol{A} \coloneqq \boldsymbol{G} + 8\ell\boldsymbol{I}_d$ denote the expected value of the augmented sum

$$\sum_{j=1}^m \boldsymbol{Y}_j + 8\ell\boldsymbol{I}_d. \tag{23}$$

Since $\boldsymbol{A} \succeq 8\ell\boldsymbol{I}_d \succ \mathbf{0}_{d\times d}$, the inverse square root $\boldsymbol{A}^{-1/2}$ is well defined. Multiplying Eq. (23) on both sides by $\boldsymbol{A}^{-1/2}$ gives

$$\begin{aligned}\boldsymbol{A}^{-1/2}\left(\sum_{j=1}^m \boldsymbol{Y}_j + 8\ell\boldsymbol{I}_d\right)\boldsymbol{A}^{-1/2} &= \sum_{j=1}^m \boldsymbol{A}^{-1/2}\boldsymbol{Y}_j\boldsymbol{A}^{-1/2} + 8\ell\boldsymbol{A}^{-1} \\ &= \sum_{j=1}^m \boldsymbol{A}^{-1/2}\boldsymbol{Y}_j\boldsymbol{A}^{-1/2} + \sum_{h=1}^{8\ell}\boldsymbol{A}^{-1}.\end{aligned}$$

Accordingly, define $\boldsymbol{Z}_j \coloneqq \boldsymbol{A}^{-1/2}\boldsymbol{Y}_j\boldsymbol{A}^{-1/2}$ for $j \in [m]$ and $\boldsymbol{Z}_{m+h} \coloneqq \boldsymbol{A}^{-1}$ for $h \in [8\ell]$. Since $\boldsymbol{G} = \sum_{j=1}^m \mathbb{E}[\boldsymbol{Y}_j]$ and $\boldsymbol{A} = \boldsymbol{G} + 8\ell\boldsymbol{I}_d$, the expected sum of these matrices satisfies

$$\begin{aligned}\mathbb{E}\left[\sum_{j=1}^{m+8\ell}\boldsymbol{Z}_j\right] &= \boldsymbol{A}^{-1/2}\left(\sum_{j=1}^m \mathbb{E}[\boldsymbol{Y}_j] + 8\ell\boldsymbol{I}_d\right)\boldsymbol{A}^{-1/2} \\ &= \boldsymbol{A}^{-1/2}(\boldsymbol{G} + 8\ell\boldsymbol{I}_d)\boldsymbol{A}^{-1/2} \\ &= \boldsymbol{A}^{-1/2}\boldsymbol{A}\boldsymbol{A}^{-1/2} = \boldsymbol{I}_d.\end{aligned}$$

We also obtain a common upper bound for every transformed summand. Since $\boldsymbol{A} \succeq 8\ell\boldsymbol{I}_d$ and matrix inversion reverses the Loewner order on positive-definite matrices, $\boldsymbol{A}^{-1} \preceq \boldsymbol{I}_d/(8\ell)$. For $j \in [m]$, the assumption $\mathbf{0}_{d\times d} \preceq \boldsymbol{Y}_j \preceq \boldsymbol{I}_d$ and the fact that congruence by the symmetric matrix $\boldsymbol{A}^{-1/2}$ preserves the Loewner order imply

$$\mathbf{0}_{d\times d} \preceq \boldsymbol{Z}_j = \boldsymbol{A}^{-1/2}\boldsymbol{Y}_j\boldsymbol{A}^{-1/2} \preceq \boldsymbol{A}^{-1} \preceq \frac{1}{8\ell}\boldsymbol{I}_d,$$

and the same bound holds for the deterministic summands because $\boldsymbol{Z}_{m+h} = \boldsymbol{A}^{-1}$. Finally, the transformed random matrices remain independent because $\boldsymbol{A}$ is deterministic, and adding deterministic matrices does not affect independence. Thus the entire collection consists of independent positive-semidefinite matrices, each bounded above by $\boldsymbol{I}_d/(8\ell)$, whose expected sum is $\boldsymbol{I}_d$.

We now apply the matrix Chernoff bound (Tropp, 2012): for independent random positive-semidefinite matrices $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n$ with $\lambda_{\max}(\boldsymbol{Z}_j) \leq R$ almost surely, $\mu_{\min} \coloneqq \lambda_{\min}(\mathbb{E}[\sum_{j=1}^n \boldsymbol{Z}_j])$, and any $\varepsilon \in (0,1)$,

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{j=1}^n \boldsymbol{Z}_j\right) \leq (1-\varepsilon)\mu_{\min}\right) \leq d\left[\frac{\exp(-\varepsilon)}{(1-\varepsilon)^{1-\varepsilon}}\right]^{\mu_{\min}/R} \leq d\exp\left(-\frac{\varepsilon^2\mu_{\min}}{2R}\right),$$

where the last inequality uses $\exp(-\varepsilon)/(1-\varepsilon)^{1-\varepsilon} \leq \exp(-\varepsilon^2/2)$ for $\varepsilon \in (0,1)$. Applying this bound with $n = m + 8\ell$, $\varepsilon = 1/2$, $\mu_{\min} = 1$, and $R = 1/(8\ell)$ gives

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{j=1}^{m+8\ell}\boldsymbol{Z}_j\right) \leq \frac{1}{2}\right) \leq d\exp\left(-\frac{(1/2)^2}{2}\cdot 8\ell\right) = d\exp(-\ell).$$

On the complementary event, which has probability at least $1 - d\exp(-\ell)$,

$$\boldsymbol{A}^{-1/2}\left(\sum_{j=1}^{m}\boldsymbol{Y}_j + 8\ell\boldsymbol{I}_d\right)\boldsymbol{A}^{-1/2} \succeq \frac{1}{2}\boldsymbol{I}_d.$$

Since congruence by $\boldsymbol{A}^{1/2}$ preserves the Loewner order, multiplying both sides on the left and right by $\boldsymbol{A}^{1/2}$ yields

$$\begin{aligned}\sum_{j=1}^{m}\boldsymbol{Y}_j + 8\ell\boldsymbol{I}_d &\succeq \frac{1}{2}\boldsymbol{A}^{1/2}\boldsymbol{I}_d\boldsymbol{A}^{1/2}\\ &= \frac{1}{2}\boldsymbol{A}\\ &= \frac{1}{2}\boldsymbol{G} + 4\ell\boldsymbol{I}_d, \end{aligned} \tag{24}$$

where the last equality uses $\boldsymbol{A} = \boldsymbol{G} + 8\ell\boldsymbol{I}_d$. Dividing Eq. (24) by $8\ell > 0$ gives

$$\begin{aligned}\boldsymbol{I}_d + \frac{1}{8\ell}\sum_{j=1}^{m}\boldsymbol{Y}_j &\succeq \frac{1}{8\ell}\left(\frac{1}{2}\boldsymbol{G} + 4\ell\boldsymbol{I}_d\right)\\ &= \frac{\boldsymbol{G}}{16\ell} + \frac{1}{2}\boldsymbol{I}_d.\end{aligned}$$

Moreover, since each $\boldsymbol{Y}_j$ is positive semidefinite and $8\ell \geq 1$,

$$\sum_{j=1}^{m}\boldsymbol{Y}_j - \frac{1}{8\ell}\sum_{j=1}^{m}\boldsymbol{Y}_j = \left(1 - \frac{1}{8\ell}\right)\sum_{j=1}^{m}\boldsymbol{Y}_j \succeq \boldsymbol{0}_{d\times d},$$

implying that $\sum_{j=1}^{m}\boldsymbol{Y}_j \succeq (1/8\ell)\sum_{j=1}^{m}\boldsymbol{Y}_j$. Combining the two inequalities yields

$$\boldsymbol{I}_d + \sum_{j=1}^{m}\boldsymbol{Y}_j \succeq \boldsymbol{I}_d + \frac{1}{8\ell}\sum_{j=1}^{m}\boldsymbol{Y}_j \succeq \frac{1}{2}\boldsymbol{I}_d + \frac{\boldsymbol{G}}{16\ell},$$

which completes the proof. □

We now present the regret bound induced by Algorithm 4.

Theorem 3 (Regret under IID time-varying arm sets with an unknown span)**.** *For fixed $p \in (1/2, 1)$ and $\delta \in (0, 1)$, let $\beta \coloneqq \sqrt{2\log(4N/\delta) + d\log(1 + T/d)} + 1$, $\ell_T \coloneqq \lceil \log(4Nd(T+1)/\delta)\rceil$, and $C_0 \coloneqq \lceil 32(2+\sqrt{2})^2 N\beta^2\ell_T\rceil$. With probability at least $1-\delta$, Algorithm 4 incurs a cumulative regret upper bounded by*

$$\mathrm{Reg}(T) \leq 2\beta\sqrt{2dC_{\mathcal{P}}T\log\left(1 + \frac{T}{d}\right)} + 8C_0^p T^{1-p} + 8C_0(N\pi_{\min})^{-\frac{1}{1-p}}\Delta_{\mathcal{P}}^{-\frac{2}{1-p}},$$

where the last term is omitted when $C_{\mathcal{P}} = 1$.

Proof of Theorem 3. If $W_{\mathcal{P}} = \{\boldsymbol{0}_d\}$, all arm features through round T are zero almost surely and $\mathrm{Reg}(T) = 0$. Following Appendix D.3, we work with the statistics formed from the original arm features with $\lambda = 1$. For $S \subseteq [N]$ and $t \in \{0, \ldots, T\}$, define the full-history statistics

$$\boldsymbol{V}_{S,t} \coloneqq \boldsymbol{I}_d + \sum_{\tau=1}^{t}\mathbf{1}\{u_\tau \in S\}\boldsymbol{x}_{\tau,a_\tau}\boldsymbol{x}_{\tau,a_\tau}^\top, \quad \boldsymbol{b}_{S,t} \coloneqq \sum_{\tau=1}^{t}\mathbf{1}\{u_\tau \in S\}y_\tau\boldsymbol{x}_{\tau,a_\tau},$$

and let $\widehat{\boldsymbol{\theta}}_{S,t} \coloneqq \boldsymbol{V}_{S,t}^{-1}\boldsymbol{b}_{S,t}$. Similarly, let $I_{\mathsf{e}}(\tau)$ indicate whether round τ is a forced-exploration round, and define the exploration statistics of each user $i \in [N]$ by

$$\boldsymbol{V}_{i,\mathsf{e},t} \coloneqq \boldsymbol{I}_d + \sum_{\tau=1}^{t}I_{\mathsf{e}}(\tau)\mathbf{1}\{u_\tau = i\}\boldsymbol{x}_{\tau,a_\tau}\boldsymbol{x}_{\tau,a_\tau}^\top, \quad \boldsymbol{b}_{i,\mathsf{e},t} \coloneqq \sum_{\tau=1}^{t}I_{\mathsf{e}}(\tau)\mathbf{1}\{u_\tau = i\}y_\tau\boldsymbol{x}_{\tau,a_\tau},$$

and let $\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t} \coloneqq \boldsymbol{V}_{i,\mathsf{e},t}^{-1}\boldsymbol{b}_{i,\mathsf{e},t}$.

Simultaneous confidence. By Lemma 12, all selected features through round T lie in $W_{\mathcal{P}}$ almost surely. Thus the observations from user i have conditional means determined by the fixed parameter $\Pi_{W_{\mathcal{P}}}\boldsymbol{\theta}_i^\star$, with $\|\Pi_{W_{\mathcal{P}}}\boldsymbol{\theta}_i^\star\| \le \|\boldsymbol{\theta}_i^\star\| \le 1$.

Fix a user i. Since the phase schedule is deterministic and u_τ and $\boldsymbol{x}_{\tau,a_\tau}$ are $\mathcal{F}_{\tau-1}$-measurable, the regressor $I_{\mathsf{e}}(\tau)\mathbf{1}\{u_\tau = i\}\boldsymbol{x}_{\tau,a_\tau}$ is $\mathcal{F}_{\tau-1}$-measurable. Applying the self-normalized concentration bound (Abbasi-yadkori et al., 2011) therefore gives, with failure probability at most $\delta/(4N)$,

$$\left\|\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t} - \Pi_{W_{\mathcal{P}}}\boldsymbol{\theta}_i^\star\right\|_{\boldsymbol{V}_{i,\mathsf{e},t}} \le \sqrt{2\log\frac{4N}{\delta} + \log\det(\boldsymbol{V}_{i,\mathsf{e},t})} + 1$$

simultaneously for all $t \in \{0,\dots,T\}$. Moreover, since $\|\boldsymbol{x}_{\tau,a_\tau}\| \le 1$ for all $\tau \in [t]$, we have $\mathrm{tr}(\boldsymbol{V}_{i,\mathsf{e},t}) \le d + n_{i,\mathsf{e},t}$, and the arithmetic-geometric mean inequality applied to the eigenvalues gives

$$\log\det(\boldsymbol{V}_{i,\mathsf{e},t}) \le d\log\left(\frac{\mathrm{tr}(\boldsymbol{V}_{i,\mathsf{e},t})}{d}\right) \le d\log\left(1 + \frac{n_{i,\mathsf{e},t}}{d}\right).$$

Consequently, by the definition of $\beta_{i,\mathsf{e},t}$ in Appendix D.2,

$$\left\|\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t} - \Pi_{W_{\mathcal{P}}}\boldsymbol{\theta}_i^\star\right\|_{\boldsymbol{V}_{i,\mathsf{e},t}} \le \beta_{i,\mathsf{e},t}. \tag{25}$$

For each population observable cluster $\mathcal{I}_c$, all observations from its members share the same projected parameter $\boldsymbol{\vartheta}_c^\star$. Since $\mathbf{1}\{u_\tau \in \mathcal{I}_c\}\boldsymbol{x}_{\tau,a_\tau}$ is also $\mathcal{F}_{\tau-1}$-measurable, applying the self-normalized concentration bound to the cluster's pooled observations gives, with failure probability at most $\delta/(4N)$,

$$\begin{aligned}\left\|\widehat{\boldsymbol{\vartheta}}_{\mathcal{I}_c,t} - \boldsymbol{\vartheta}_c^\star\right\|_{\boldsymbol{V}_{\mathcal{I}_c,t}} &\le \sqrt{2\log\frac{4N}{\delta} + \log\det(\boldsymbol{V}_{\mathcal{I}_c,t})} + 1 \\ &\le \sqrt{2\log\frac{4N}{\delta} + d\log\left(1 + \frac{T}{d}\right)} + 1\end{aligned} \tag{26}$$

simultaneously for all $t \in \{0,\dots,T\}$. Here the second inequality uses the fact that at most T observations contribute to any cluster's design matrix.

There are N user-level confidence events and $C_{\mathcal{P}} \le N$ cluster-level confidence events. A union bound therefore shows that both bounds hold simultaneously for all users, clusters, and times with failure probability at most

$$(N + C_{\mathcal{P}})\frac{\delta}{4N} \le \frac{\delta}{2}.$$

Joint arrival and feature concentration. Fix a phase s whose graph update is executed by time T, that is, $t_s + 1 \le T$; the update at round $t_s + 1$ uses the exploration statistics through round t_s. Let $\mathcal{E}_s \coloneqq \{h_s + 1, \dots, t_s\}$ denote the M_s forced-exploration rounds of phase s.

For each user i and every forced-exploration round t, define

$$\boldsymbol{Y}_{i,t} \coloneqq \mathbf{1}\{u_t = i\}\boldsymbol{x}_{t,a_t}\boldsymbol{x}_{t,a_t}^\top.$$

For fixed i, the IID generation of the arm sets, their independence of the user arrivals, the IID arrivals, and the fresh independent uniform sampling of arm indices imply that these matrices are independent across $t \in \mathcal{E}_s$. Since every arm feature has norm at most one,

$$\mathbf{0}_{d\times d} \preceq \boldsymbol{Y}_{i,t} \preceq \boldsymbol{I}_d.$$

Moreover, the exploration index a_t is independent of $(u_t, \mathcal{A}_t)$, and u_t is independent of $\mathcal{A}_t$, so

$$\begin{aligned}\mathbb{E}[\boldsymbol{Y}_{i,t}] &= \sum_{k=1}^{K}\mathbb{P}(a_t = k)\mathbb{E}\left[\mathbf{1}\{u_t = i\}\boldsymbol{x}_{t,k}\boldsymbol{x}_{t,k}^\top\right] \\ &= \frac{\pi_i}{K}\sum_{k=1}^{K}\mathbb{E}[\boldsymbol{x}_{t,k}\boldsymbol{x}_{t,k}^\top] \\ &= \pi_i\boldsymbol{\Sigma}_{\mathcal{P}}.\end{aligned}$$

Consequently, $\sum_{t\in\mathcal{E}_s}\mathbb{E}[\boldsymbol{Y}_{i,t}] = M_s\pi_i\boldsymbol{\Sigma}_{\mathcal{P}}$.

The exploration design at round t_s includes observations from both phase s and all preceding phases. Applying Lemma 13 with $\ell = \ell_T$ to $\{\boldsymbol{Y}_{i,\tau}\}_{\tau\in\mathcal{E}_s}$, we obtain, with probability at least $1 - d\exp(-\ell_T)$,

$$\begin{aligned}\boldsymbol{V}_{i,\mathsf{e},t_s} &= \boldsymbol{I}_d + \sum_{r=0}^{s-1}\sum_{\tau\in\mathcal{E}_r}\boldsymbol{Y}_{i,\tau} + \sum_{\tau\in\mathcal{E}_s}\boldsymbol{Y}_{i,\tau}\\ &\succeq \boldsymbol{I}_d + \sum_{\tau\in\mathcal{E}_s}\boldsymbol{Y}_{i,\tau}\\ &\succeq \frac{1}{2}\boldsymbol{I}_d + \frac{M_s\pi_i}{16\ell_T}\boldsymbol{\Sigma}_{\mathcal{P}},\end{aligned} \tag{27}$$

where the first inequality holds since $\sum_{r=0}^{s-1}\sum_{\tau\in\mathcal{E}_r}\boldsymbol{Y}_{i,\tau} \succeq \boldsymbol{0}_{d\times d}$, and the second follows from Lemma 13.

Taking a union bound over at most $N(T+1)$ user-phase pairs with graph updates executed by time T, Eq. (27) holds simultaneously for all such pairs with failure probability at most

$$N(T+1)d\exp(-\ell_T) \le \frac{\delta}{4}.$$

Combining this event with the preceding simultaneous confidence event and the probability-one span event gives an intersection with probability at least

$$1 - \frac{\delta}{2} - \frac{\delta}{4} = 1 - \frac{3\delta}{4} \ge 1 - \delta.$$

The remainder of the proof is carried out on this intersection.

No false splits and edge precision. By Appendix D.3, the edge separation test computed from the reduced-coordinate statistics has the same test statistic and threshold as the test computed from the original arm features, so it suffices to analyze the latter.

If $i, j \in \mathcal{I}_c$, their common projected parameter satisfies $\boldsymbol{\vartheta}_c^\star = \Pi_{W_{\mathcal{P}}}\boldsymbol{\theta}_i^\star = \Pi_{W_{\mathcal{P}}}\boldsymbol{\theta}_j^\star$. Applying Lemma 4 with $\boldsymbol{C} = \boldsymbol{V}_{i,\mathsf{e},t_s}$ and $\boldsymbol{D} = \boldsymbol{V}_{j,\mathsf{e},t_s}$ and evaluating the objective at $\boldsymbol{z} = \boldsymbol{\vartheta}_c^\star$ yields

$$\begin{aligned}&\left\|\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s} - \widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s}\right\|^2_{\boldsymbol{H}_{ij,t_s}}\\ &\quad\le \left\|\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s} - \boldsymbol{\vartheta}_c^\star\right\|^2_{\boldsymbol{V}_{i,\mathsf{e},t_s}} + \left\|\widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s} - \boldsymbol{\vartheta}_c^\star\right\|^2_{\boldsymbol{V}_{j,\mathsf{e},t_s}}\\ &\quad\le \beta^2_{i,\mathsf{e},t_s} + \beta^2_{j,\mathsf{e},t_s},\end{aligned}$$

where the last inequality follows from Eq. (25). Since Eq. (20) deletes an edge only when the test statistic strictly exceeds the threshold, no edge within a cluster is deleted.

We next bound the difference between users whose edge survives a graph update. Define $\boldsymbol{D}_s := (1/2)\boldsymbol{I}_d + (M_s\pi_{\min}/16\ell_T)\boldsymbol{\Sigma}_{\mathcal{P}}$. By Eq. (27) and $\pi_i, \pi_j \ge \pi_{\min}$,

$$\boldsymbol{V}_{i,\mathsf{e},t_s} \succeq \boldsymbol{D}_s \succ \boldsymbol{0}_{d\times d}, \qquad \boldsymbol{V}_{j,\mathsf{e},t_s} \succeq \boldsymbol{D}_s \succ \boldsymbol{0}_{d\times d}.$$

Since matrix inversion reverses the Loewner order on positive-definite matrices,

$$\boldsymbol{V}^{-1}_{i,\mathsf{e},t_s} \preceq \boldsymbol{D}_s^{-1}, \qquad \boldsymbol{V}^{-1}_{j,\mathsf{e},t_s} \preceq \boldsymbol{D}_s^{-1},$$

and hence $\boldsymbol{V}^{-1}_{i,\mathsf{e},t_s} + \boldsymbol{V}^{-1}_{j,\mathsf{e},t_s} \preceq 2\boldsymbol{D}_s^{-1}$. Inverting both sides and reversing the order once more gives

$$\begin{aligned}\boldsymbol{H}_{ij,t_s} &\succeq (2\boldsymbol{D}_s^{-1})^{-1}\\ &= \frac{1}{2}\boldsymbol{D}_s\\ &= \frac{1}{4}\boldsymbol{I}_d + \frac{M_s\pi_{\min}}{32\ell_T}\boldsymbol{\Sigma}_{\mathcal{P}}\\ &\succeq \frac{M_s\pi_{\min}}{32\ell_T}\boldsymbol{\Sigma}_{\mathcal{P}}.\end{aligned} \tag{28}$$

For an edge (i, j) surviving the graph update of phase s, the deletion rule gives

$$\left\|\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s} - \widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s}\right\|_{\boldsymbol{H}_{ij,t_s}} \le \sqrt{\beta^2_{i,\mathsf{e},t_s} + \beta^2_{j,\mathsf{e},t_s}}.$$

Moreover, $\boldsymbol{V}^{-1}_{j,\mathsf{e},t_s} \succeq \mathbf{0}_{d\times d}$ implies $\boldsymbol{V}^{-1}_{i,\mathsf{e},t_s} + \boldsymbol{V}^{-1}_{j,\mathsf{e},t_s} \succeq \boldsymbol{V}^{-1}_{i,\mathsf{e},t_s}$, so order reversal gives $\boldsymbol{H}_{ij,t_s} \preceq \boldsymbol{V}_{i,\mathsf{e},t_s}$, and similarly $\boldsymbol{H}_{ij,t_s} \preceq \boldsymbol{V}_{j,\mathsf{e},t_s}$. Together with Eq. (25), we obtain

$$\left\|\Pi_{W_{\mathcal{P}}}\boldsymbol{\theta}^\star_i - \widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s}\right\|_{\boldsymbol{H}_{ij,t_s}} \le \left\|\Pi_{W_{\mathcal{P}}}\boldsymbol{\theta}^\star_i - \widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s}\right\|_{\boldsymbol{V}_{i,\mathsf{e},t_s}} \le \beta_{i,\mathsf{e},t_s},$$

with the analogous bound for user j. Decomposing the gap between the two projected parameters through their estimators and applying the triangle inequality gives

$$\begin{aligned}
&\left\|\Pi_{W_{\mathcal{P}}}(\boldsymbol{\theta}^\star_i - \boldsymbol{\theta}^\star_j)\right\|_{\boldsymbol{H}_{ij,t_s}} \\
&\le \left\|\Pi_{W_{\mathcal{P}}}\boldsymbol{\theta}^\star_i - \widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s}\right\|_{\boldsymbol{H}_{ij,t_s}} + \left\|\widehat{\boldsymbol{\theta}}_{i,\mathsf{e},t_s} - \widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s}\right\|_{\boldsymbol{H}_{ij,t_s}} + \left\|\widehat{\boldsymbol{\theta}}_{j,\mathsf{e},t_s} - \Pi_{W_{\mathcal{P}}}\boldsymbol{\theta}^\star_j\right\|_{\boldsymbol{H}_{ij,t_s}} \\
&\le \beta_{i,\mathsf{e},t_s} + \sqrt{\beta^2_{i,\mathsf{e},t_s} + \beta^2_{j,\mathsf{e},t_s}} + \beta_{j,\mathsf{e},t_s} \\
&\le (2+\sqrt{2})\beta,
\end{aligned}$$

where the last inequality uses $\beta_{i,\mathsf{e},t_s}, \beta_{j,\mathsf{e},t_s} \le \beta$, which holds because $n_{i,\mathsf{e},t_s}, n_{j,\mathsf{e},t_s} \le T$.

Since $\boldsymbol{\Sigma}_{\mathcal{P}}$ is symmetric with range $W_{\mathcal{P}}$, $\Pi_{W_{\mathcal{P}}}\boldsymbol{\Sigma}_{\mathcal{P}} = \boldsymbol{\Sigma}_{\mathcal{P}}\Pi_{W_{\mathcal{P}}} = \boldsymbol{\Sigma}_{\mathcal{P}}$. Consequently, for every $\boldsymbol{u} \in \mathbb{R}^d$,

$$\|\Pi_{W_{\mathcal{P}}}\boldsymbol{u}\|^2_{\boldsymbol{\Sigma}_{\mathcal{P}}} = \boldsymbol{u}^\top \Pi_{W_{\mathcal{P}}}\boldsymbol{\Sigma}_{\mathcal{P}}\Pi_{W_{\mathcal{P}}}\boldsymbol{u} = \boldsymbol{u}^\top\boldsymbol{\Sigma}_{\mathcal{P}}\boldsymbol{u} = \|\boldsymbol{u}\|^2_{\boldsymbol{\Sigma}_{\mathcal{P}}}. \tag{29}$$

Combining Eq. (29) with Eq. (28) yields

$$\sqrt{\frac{M_s\pi_{\min}}{32\ell_T}}\left\|\boldsymbol{\theta}^\star_i - \boldsymbol{\theta}^\star_j\right\|_{\boldsymbol{\Sigma}_{\mathcal{P}}} \le \left\|\Pi_{W_{\mathcal{P}}}(\boldsymbol{\theta}^\star_i - \boldsymbol{\theta}^\star_j)\right\|_{\boldsymbol{H}_{ij,t_s}} \le (2+\sqrt{2})\beta.$$

Therefore, using $M_s = C_0 2^s$ and $C_0 \ge 32(2+\sqrt{2})^2 N\beta^2\ell_T$, we obtain

$$\begin{aligned}
\left\|\boldsymbol{\theta}^\star_i - \boldsymbol{\theta}^\star_j\right\|_{\boldsymbol{\Sigma}_{\mathcal{P}}} &\le \frac{\sqrt{32\ell_T}(2+\sqrt{2})\beta}{\sqrt{M_s\pi_{\min}}} \\
&= \frac{\sqrt{32\ell_T}(2+\sqrt{2})\beta}{\sqrt{C_0 2^s\pi_{\min}}} \\
&\le \frac{2^{-s/2}}{\sqrt{N\pi_{\min}}}.
\end{aligned} \tag{30}$$

Eq. (30) bounds the reward-relevant parameter difference across every edge that survives the graph update of phase s.

Recovery and pre-recovery regret. If $C_{\mathcal{P}} = 1$, no within-cluster edge is deleted, so the graph remains complete and every adaptive round uses the correct all-user pool; in this case, we set $s_\star := 0$. For $C_{\mathcal{P}} \ge 2$, define

$$s_\star := \min\left\{s \in \mathbb{Z}_{\ge 0} : \frac{2^{-s/2}}{\sqrt{N\pi_{\min}}} < \Delta_{\mathcal{P}}\right\}.$$

This index is finite because $\Delta_{\mathcal{P}} > 0$ and $\pi_{\min} > 0$, but the algorithm need not reach this phase by time T. The index $s_\star$ is used only in the analysis and is not an input to the algorithm.

Suppose that $C_{\mathcal{P}} \ge 2$ and that an edge (i, j) between different population observable clusters survives an executed graph update in a phase $s \ge s_\star$. By the definition of $\Delta_{\mathcal{P}}$ and Eq. (30),

$$\Delta_{\mathcal{P}} \le \left\|\boldsymbol{\theta}^\star_i - \boldsymbol{\theta}^\star_j\right\|_{\boldsymbol{\Sigma}_{\mathcal{P}}} \le \frac{2^{-s/2}}{\sqrt{N\pi_{\min}}} \le \frac{2^{-s_\star/2}}{\sqrt{N\pi_{\min}}} < \Delta_{\mathcal{P}},$$

a contradiction. Thus every executed graph update in a phase $s \ge s_\star$ removes all remaining cross-cluster edges. Since the graph starts complete and no within-cluster edge is deleted, each self-inclusive direct neighborhood then equals the corresponding population observable cluster.

Let $\mathrm{Reg}_{\mathrm{pre}}(T)$ denote the regret incurred on adaptive-selection rounds in phases $s < s_\star$ that are executed by time T. Forced-exploration rounds are accounted for separately below. If $s_\star > 0$, the defining inequality fails at $s_\star - 1$, so

$$\frac{2^{-(s_\star-1)/2}}{\sqrt{N\pi_{\min}}} \geq \Delta_{\mathcal{P}},$$

and hence,

$$2^{s_\star-1} \leq \frac{1}{N\pi_{\min}\Delta_{\mathcal{P}}^2}.$$

Since $C_0 \geq 1$, $L_s = \lceil C_0 2^{s/(1-p)} \rceil \leq 2C_0 2^{s/(1-p)}$. Each round contributes at most 2 to regret, giving

$$\begin{aligned}
\mathrm{Reg}_{\mathrm{pre}}(T) &\leq 2\sum_{s=0}^{s_\star-1} L_s \\
&\leq 4C_0 \sum_{s=0}^{s_\star-1} 2^{\frac{s}{1-p}} \\
&= 4C_0 2^{\frac{s_\star-1}{1-p}} \sum_{h=0}^{s_\star-1} 2^{-\frac{h}{1-p}} \\
&\leq \frac{4C_0 2^{\frac{s_\star-1}{1-p}}}{1-2^{-\frac{1}{1-p}}} \\
&\leq 8C_0 (N\pi_{\min})^{-\frac{1}{1-p}} \Delta_{\mathcal{P}}^{-\frac{2}{1-p}}.
\end{aligned}$$

The last step uses $1 - 2^{-1/(1-p)} \geq 1/2$ for any $p \in (0,1)$ and the preceding bound on $2^{s_\star-1}$. Even if the horizon ends before recovery or during an adaptive subphase, the same bound holds because the sum of full subphase lengths only overcounts executed rounds. If $s_\star = 0$, including the case $C_{\mathcal{P}} = 1$, there are no pre-recovery adaptive rounds and $\mathrm{Reg}_{\mathrm{pre}}(T) = 0$.

Post-recovery regret. Let $\mathcal{T}_{\mathrm{post}} \subseteq [T]$ contain the adaptive-selection rounds in phases $s \geq s_\star$. For every $t \in [T]$, let c_t label the population observable cluster of u_t. On $\mathcal{T}_{\mathrm{post}}$, the neighborhood pool coincides with $\mathcal{I}_{c_t}$. The corresponding estimator uses all observations previously collected from users in that cluster, including observations collected before recovery.

By Eq. (26) and the Cauchy-Schwarz inequality, every available arm satisfies

$$\left| \boldsymbol{x}_{t,k}^\top \left(\widehat{\boldsymbol{\theta}}_{\mathcal{I}_{c_t},t-1} - \boldsymbol{\vartheta}_{c_t}^\star \right) \right| \leq \beta \|\boldsymbol{x}_{t,k}\|_{\boldsymbol{V}_{\mathcal{I}_{c_t},t-1}^{-1}}.$$

Since $\boldsymbol{x}_{t,k}^\top \boldsymbol{\theta}_{u_t}^\star = \boldsymbol{x}_{t,k}^\top \boldsymbol{\vartheta}_{c_t}^\star$ for every $k \in [K]$ almost surely, the optimal index $a_t^\star$ defined in Appendix D.1 satisfies $a_t^\star \in \mathrm{argmax}_{k\in[K]} \boldsymbol{x}_{t,k}^\top \boldsymbol{\vartheta}_{c_t}^\star$. For every $t \in \mathcal{T}_{\mathrm{post}}$,

$$\begin{aligned}
\boldsymbol{x}_{t,a_t^\star}^\top \boldsymbol{\vartheta}_{c_t}^\star &\leq \boldsymbol{x}_{t,a_t^\star}^\top \widehat{\boldsymbol{\theta}}_{\mathcal{I}_{c_t},t-1} + \beta \|\boldsymbol{x}_{t,a_t^\star}\|_{\boldsymbol{V}_{\mathcal{I}_{c_t},t-1}^{-1}} \\
&\leq \boldsymbol{x}_{t,a_t}^\top \widehat{\boldsymbol{\theta}}_{\mathcal{I}_{c_t},t-1} + \beta \|\boldsymbol{x}_{t,a_t}\|_{\boldsymbol{V}_{\mathcal{I}_{c_t},t-1}^{-1}} \\
&\leq \boldsymbol{x}_{t,a_t}^\top \boldsymbol{\vartheta}_{c_t}^\star + 2\beta \|\boldsymbol{x}_{t,a_t}\|_{\boldsymbol{V}_{\mathcal{I}_{c_t},t-1}^{-1}}.
\end{aligned}$$

The instantaneous regret is therefore at most $2\beta \|\boldsymbol{x}_{t,a_t}\|_{\boldsymbol{V}_{\mathcal{I}_{c_t},t-1}^{-1}}$.

To sum these widths, fix a population observable cluster $c \in [C_{\mathcal{P}}]$, let $\mathcal{T}_c := \{t \in [T] : u_t \in \mathcal{I}_c\}$, and apply the elliptical potential lemma to the arms $(\boldsymbol{x}_{t,a_t})_{t\in\mathcal{T}_c}$, whose design starts from $\boldsymbol{I}_d$ and whose features satisfy $\|\boldsymbol{x}_{t,a_t}\| \leq 1$. Since the set of post-recovery rounds of cluster c is a subset of $\mathcal{T}_c$ and all squared widths are nonnegative,

$$\sum_{t\in\mathcal{T}_{\mathrm{post}}, c_t=c} \|\boldsymbol{x}_{t,a_t}\|^2_{\boldsymbol{V}_{\mathcal{I}_c,t-1}^{-1}} \leq 2\log\det(\boldsymbol{V}_{\mathcal{I}_c,T}) \leq 2d\log\left(1+\frac{T}{d}\right),$$

where the last inequality follows from the determinant-trace inequality and $\mathrm{tr}(\boldsymbol{V}_{\mathcal{I}_c,T}) \le d+T$. Summing over the $C_{\mathcal{P}}$ clusters gives

$$\sum_{t\in\mathcal{T}_{\mathrm{post}}} \|\boldsymbol{x}_{t,a_t}\|^2_{\boldsymbol{V}^{-1}_{\mathcal{I}_{c_t},t-1}} \le 2dC_{\mathcal{P}} \log\left(1+\frac{T}{d}\right).$$

Finally, the Cauchy-Schwarz inequality and $|\mathcal{T}_{\mathrm{post}}| \le T$ give

$$\begin{aligned}\mathrm{Reg}_{\mathrm{post}}(T) &\le 2\beta \sum_{t\in\mathcal{T}_{\mathrm{post}}} \|\boldsymbol{x}_{t,a_t}\|_{\boldsymbol{V}^{-1}_{\mathcal{I}_{c_t},t-1}} \\ &\le 2\beta \sqrt{|\mathcal{T}_{\mathrm{post}}| \sum_{t\in\mathcal{T}_{\mathrm{post}}} \|\boldsymbol{x}_{t,a_t}\|^2_{\boldsymbol{V}^{-1}_{\mathcal{I}_{c_t},t-1}}} \\ &\le 2\beta \sqrt{2dC_{\mathcal{P}}T\log\left(1+\frac{T}{d}\right)}.\end{aligned}$$

Forced-exploration regret. Let $E(T)$ be the number of forced-exploration rounds executed by time T, and let $\bar{s}$ be the last phase started by that time. If $\bar{s}=0$, at most C_0 exploration rounds can have occurred, so

$$E(T) \le \min\{T, C_0\} \le C_0^p T^{1-p}.$$

The second inequality follows by considering $T \le C_0$ and $T \ge C_0$ separately.

If $\bar{s} \ge 1$, starting phase $\bar{s}$ requires completing the adaptive subphase of phase $\bar{s}-1$. Hence

$$C_0 2^{\frac{\bar{s}-1}{1-p}} \le L_{\bar{s}-1} \le T,$$

thus,

$$2^{\bar{s}-1} \le \left(\frac{T}{C_0}\right)^{1-p}.$$

Even if phase $\bar{s}$ is incomplete, its executed exploration rounds are at most $M_{\bar{s}}$. Using $M_s = C_0 2^s$ therefore gives

$$\begin{aligned}E(T) &\le \sum_{s=0}^{\bar{s}} M_s = C_0(2^{\bar{s}+1}-1) \\ &< 4C_0 2^{\bar{s}-1} \\ &\le 4C_0\left(\frac{T}{C_0}\right)^{1-p} = 4C_0^p T^{1-p}.\end{aligned}$$

The bound $E(T) \le 4C_0^p T^{1-p}$ thus holds in both cases. Since each exploration round incurs regret at most 2,

$$\mathrm{Reg}_{\mathrm{explore}}(T) \le 2E(T) \le 8C_0^p T^{1-p}.$$

Every round belongs to exactly one of the three categories considered above. For $C_{\mathcal{P}} \ge 2$, summing their contributions yields

$$\begin{aligned}\mathrm{Reg}(T) &= \mathrm{Reg}_{\mathrm{pre}}(T) + \mathrm{Reg}_{\mathrm{post}}(T) + \mathrm{Reg}_{\mathrm{explore}}(T) \\ &\le 2\beta\sqrt{2dC_{\mathcal{P}}T\log\left(1+\frac{T}{d}\right)} + 8C_0^p T^{1-p} + 8C_0(N\pi_{\min})^{-\frac{1}{1-p}}\Delta_{\mathcal{P}}^{-\frac{2}{1-p}}.\end{aligned}$$

If $C_{\mathcal{P}}=1$, then $\mathrm{Reg}_{\mathrm{pre}}(T)=0$, and the same bound holds without the last term. Finally, since $\beta^2 = \widetilde{\mathcal{O}}(d)$ and $\ell_T = \mathcal{O}(\log(NdT/\delta))$, we have $C_0 = \widetilde{\mathcal{O}}(Nd)$; hence, for any fixed $p \in (1/2,1)$, the exploration term $\widetilde{\mathcal{O}}((Nd)^p T^{1-p})$ grows strictly slower than $\sqrt{T}$, and the term $\widetilde{\mathcal{O}}(d\sqrt{C_{\mathcal{P}}T})$ asymptotically dominates. □

E Implementation Details and Computational Analysis

Algorithm 1 forms the pooled statistics of user i in phase s by adding up the per-user statistics of all users in the neighborhood,

$$\boldsymbol{V}_{\mathcal{N}_s(i),t-1} = \lambda \boldsymbol{I}_d + \sum_{j \in \mathcal{N}_s(i)} \left(\boldsymbol{V}_{j,t-1} - \lambda \boldsymbol{I}_d\right), \qquad \boldsymbol{b}_{\mathcal{N}_s(i),t-1} = \sum_{j \in \mathcal{N}_s(i)} \boldsymbol{b}_{j,t-1}.$$

Computing these sums literally requires adding up to N matrices of size $d \times d$ in every round. We show that the same statistics can be computed exactly at a much lower cost, without adding matrices over a neighborhood. The implementation relies on two simple observations. First, the arm set is fixed, so the design matrix of each user is constructed from the same K outer products $(\boldsymbol{x}_1\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_K\boldsymbol{x}_K^\top)$. Consequently, the design matrix is described by the number of times each of the K arms has been selected. Second, every arm lies in the B-dimensional subspace $W = \mathrm{span}(\mathcal{A})$, so each arm can be represented as a linear combination of B basis vectors of W.

To use the second observation, define $\boldsymbol{Q} \in \mathbb{R}^{d \times B}$ as a matrix whose columns constitute an orthonormal basis for W. This basis can be constructed, for instance, by applying the Gram-Schmidt process to the basis arms in $\mathcal{B}$. Each arm $\boldsymbol{x}_k$ is then represented by its coordinates $\boldsymbol{z}_k \coloneqq \boldsymbol{Q}^\top \boldsymbol{x}_k \in \mathbb{R}^B$, which satisfy $\boldsymbol{x}_k = \boldsymbol{Q}\boldsymbol{z}_k$, and we write $\boldsymbol{Z} \in \mathbb{R}^{K \times B}$ for the matrix whose k-th row is $\boldsymbol{z}_k^\top$. Since W is spanned by the K arms, $B \le \min\{d, K\}$. Representing the statistics in these coordinates entails no loss of information relevant to the algorithm. For any statistics $(\boldsymbol{V}, \boldsymbol{b})$ constructed from the selected arms and the observed rewards with ridge term $\lambda \boldsymbol{I}_d$, let $(\bar{\boldsymbol{V}}, \bar{\boldsymbol{b}})$ denote the statistics constructed from the corresponding observations in the coordinates of W with ridge term $\lambda \boldsymbol{I}_B$; then

$$\boldsymbol{V}^{-1} = \boldsymbol{Q}\bar{\boldsymbol{V}}^{-1}\boldsymbol{Q}^\top + \lambda^{-1}\left(\boldsymbol{I}_d - \boldsymbol{Q}\boldsymbol{Q}^\top\right), \qquad \boldsymbol{b} = \boldsymbol{Q}\bar{\boldsymbol{b}},$$

so the estimated reward $\boldsymbol{x}_k^\top \boldsymbol{V}^{-1}\boldsymbol{b} = \boldsymbol{z}_k^\top \bar{\boldsymbol{V}}^{-1}\bar{\boldsymbol{b}}$ and the confidence width $\|\boldsymbol{x}_k\|_{\boldsymbol{V}^{-1}} = \|\boldsymbol{z}_k\|_{\bar{\boldsymbol{V}}^{-1}}$ are unchanged for all arms $k \in [K]$. The same holds for the harmonic-mean statistic, because the difference of two exploration estimates also lies in W (see Appendix D.3 for the analogous identities). Hence the UCB rule in Line 25 of Algorithm 1 and the separation test in Line 8 of Algorithm 2 produce the same outcomes when computed in these coordinates.

Scalar propagation of the main statistics. Because the arm set is fixed, the design matrix of a user is determined by how many times each arm has been pulled, and the corresponding reward vector by the sum of the rewards observed on each arm. Formally, let $n_{i,k}$ denote the number of times user i has pulled arm k so far, and let $Y_{i,k}$ denote the sum of the corresponding rewards. Henceforth, counts, reward sums, statistics, and neighborhoods refer to their current values, and we suppress the round index for simplicity; the statistics held at the beginning of round t correspond to the statistics indexed by $t-1$ in Algorithm 1.

In the coordinates of W, the main statistics of user i are then

$$\bar{\boldsymbol{V}}_i = \lambda \boldsymbol{I}_B + \sum_{k=1}^{K} n_{i,k}\boldsymbol{z}_k\boldsymbol{z}_k^\top, \qquad \bar{\boldsymbol{b}}_i = \sum_{k=1}^{K} Y_{i,k}\boldsymbol{z}_k.$$

Let $\mathcal{N}(i)$ denote the neighborhood of user i, including user i itself, in the most recently updated graph; that is, $\mathcal{N}(i) = [N]$ before the first graph update, and $\mathcal{N}(i) = \mathcal{N}_s(i)$ from the graph update of phase s until the graph update of phase $s+1$. Since a pooled statistic is the sum of the per-user statistics in the neighborhood with a single ridge term, the same representation holds with the pooled counts $n_{\mathcal{N}(i),k} \coloneqq \sum_{j \in \mathcal{N}(i)} n_{j,k}$ and the pooled reward sums $Y_{\mathcal{N}(i),k} \coloneqq \sum_{j \in \mathcal{N}(i)} Y_{j,k}$:

$$\bar{\boldsymbol{V}}_{\mathcal{N}(i)} = \lambda \boldsymbol{I}_B + \sum_{k=1}^{K} n_{\mathcal{N}(i),k}\boldsymbol{z}_k\boldsymbol{z}_k^\top, \qquad \bar{\boldsymbol{b}}_{\mathcal{N}(i)} = \sum_{k=1}^{K} Y_{\mathcal{N}(i),k}\boldsymbol{z}_k. \tag{31}$$

For each user i and each arm k, the implementation stores only four values:

- $n_{i,k} = \sum_\tau \mathbf{1}\{u_\tau = i, a_\tau = k\}$: the number of rounds in which arm k is selected by user i.
- $Y_{i,k} = \sum_\tau \mathbf{1}\{u_\tau = i, a_\tau = k\} y_\tau$: the sum of the rewards observed in these rounds.

- $n_{\mathcal{N}(i),k} = \sum_{j \in \mathcal{N}(i)} n_{j,k}$: the number of rounds in which arm k is selected by any user in $\mathcal{N}(i)$.
- $Y_{\mathcal{N}(i),k} = \sum_{j \in \mathcal{N}(i)} Y_{j,k}$: the sum of the rewards observed in these rounds.

The implementation does not store the pooled design matrix $\bar{\boldsymbol{V}}_{\mathcal{N}(i)}$ or the pooled reward vector $\bar{\boldsymbol{b}}_{\mathcal{N}(i)}$ of any user. The pooled statistics are needed only in adaptive rounds, to compute the UCB scores of the served user u_t, and in such a round the implementation builds the pooled statistics of u_t alone from the stored values $n_{\mathcal{N}(u_t),k}$ and $Y_{\mathcal{N}(u_t),k}$, using Eq. (31) with $i = u_t$. Forced-exploration rounds select arms in the fixed round-robin order and therefore do not require the pooled statistics.

When user u_t pulls arm a_t and observes y_t, the observation enters the pooled statistics of exactly the users in $\mathcal{N}(u_t)$. Moreover, the observation changes only the counts and reward sums associated with a_t. Hence, instead of adding the matrix $\boldsymbol{z}_{a_t}\boldsymbol{z}_{a_t}^\top$ and the vector $y_t\boldsymbol{z}_{a_t}$ to the pooled statistics of every user in $\mathcal{N}(u_t)$, the implementation increases $n_{\mathcal{N}(i),a_t}$ by one and adds y_t to $Y_{\mathcal{N}(i),a_t}$ for every $i \in \mathcal{N}(u_t)$, and updates n_{u_t,a_t} and Y_{u_t,a_t} in the same way. We refer to this update as scalar propagation. Scalar propagation costs $\mathcal{O}(|\mathcal{N}(u_t)|) \le \mathcal{O}(N)$ operations per round, compared with $\mathcal{O}(|\mathcal{N}(u_t)|B^2)$ for adding the matrices, and is the only per-round operation whose cost grows with N.

Pooled UCB without matrix accumulation. During each adaptive-selection round, the UCB rule for the served user u_t requires the pooled design matrix $\bar{\boldsymbol{V}} \coloneqq \bar{\boldsymbol{V}}_{\mathcal{N}(u_t)}$ and the pooled reward vector $\bar{\boldsymbol{b}} \coloneqq \bar{\boldsymbol{b}}_{\mathcal{N}(u_t)}$. These quantities are computed on demand using the stored pooled counts and reward sums as follows:

$$\bar{\boldsymbol{V}} = \lambda \boldsymbol{I}_B + \sum_{k=1}^{K} n_{\mathcal{N}(u_t),k} \boldsymbol{z}_k \boldsymbol{z}_k^\top,$$

and

$$\bar{\boldsymbol{b}} = \sum_{k=1}^{K} Y_{\mathcal{N}(u_t),k} \boldsymbol{z}_k.$$

Constructing $\bar{\boldsymbol{V}}$ requires adding K matrices of size $B \times B$, which takes $\mathcal{O}(KB^2)$ operations regardless of the neighborhood size, $|\mathcal{N}(u_t)|$. The UCB score of each arm k consists of the estimated reward $\boldsymbol{z}_k^\top \bar{\boldsymbol{V}}^{-1}\bar{\boldsymbol{b}}$ and the confidence width $\|\boldsymbol{z}_k\|_{\bar{\boldsymbol{V}}^{-1}}$, so the implementation needs products with $\bar{\boldsymbol{V}}^{-1}$ but never the inverse matrix itself. Such products are computed through the Cholesky factorization $\bar{\boldsymbol{V}} = \boldsymbol{L}\boldsymbol{L}^\top$ with a lower-triangular matrix $\boldsymbol{L}$: the factorization costs $\mathcal{O}(B^3)$ operations, the estimate $\bar{\boldsymbol{V}}^{-1}\bar{\boldsymbol{b}}$ then follows from two triangular solves in $\mathcal{O}(B^2)$ operations, and each width follows from $\|\boldsymbol{z}_k\|_{\bar{\boldsymbol{V}}^{-1}} = \|\boldsymbol{L}^{-1}\boldsymbol{z}_k\|$ by one triangular solve, for a total of $\mathcal{O}(KB^2)$ operations over the K arms.

Since $B \le K$, forming $\bar{\boldsymbol{V}}$ dominates the factorization, and an adaptive round costs $\mathcal{O}(KB^2)$ operations in addition to scalar propagation. A forced-exploration round instead selects the next basis arm from the pointer of the served user in $\mathcal{O}(1)$ operations. By Lemma 1, the exploration design matrix is determined by the committed count, so committing a completed block updates only the count and the exploration reward vector in $\mathcal{O}(B^2)$ operations, that is, $\mathcal{O}(B)$ operations per exploration round when amortized over the B rounds of the block.

Graph updates. The separation test benefits from an additional structural fact. By Lemma 1, the exploration design matrix of every user i equals $\lambda \boldsymbol{I}_d + n_{i,\mathrm{e}}\boldsymbol{\Sigma}_{\mathcal{B}}$, so the matrices of different users differ only through the scalar committed count $n_{i,\mathrm{e}}$. In the coordinates of W, let $\bar{\boldsymbol{\Sigma}}_{\mathcal{B}} \coloneqq \boldsymbol{Q}^\top \boldsymbol{\Sigma}_{\mathcal{B}} \boldsymbol{Q} = \boldsymbol{U}\,\mathrm{diag}(\zeta_1, \ldots, \zeta_B)\boldsymbol{U}^\top$ be an eigendecomposition with an orthogonal matrix $\boldsymbol{U}$, which is computed once in preprocessing. Note that $\zeta_1, \ldots, \zeta_B$ denote the eigenvalues of $\bar{\boldsymbol{\Sigma}}_{\mathcal{B}}$. Every exploration design matrix then has the same eigenvectors,

$$\lambda \boldsymbol{I}_B + n_{i,\mathrm{e}}\bar{\boldsymbol{\Sigma}}_{\mathcal{B}} = \boldsymbol{U}\,\mathrm{diag}\left(\lambda + n_{i,\mathrm{e}}\zeta_1, \ldots, \lambda + n_{i,\mathrm{e}}\zeta_B\right)\boldsymbol{U}^\top,$$

so after the change of basis given by $\boldsymbol{U}$, every exploration design matrix becomes diagonal. Diagonal matrices are easy to handle: because a diagonal matrix is inverted entrywise, the matrices used in

the separation test are diagonal in the same basis. Writing $\bar{\boldsymbol{V}}_{i,\mathsf{e}} \coloneqq \lambda \boldsymbol{I}_B + n_{i,\mathsf{e}} \bar{\boldsymbol{\Sigma}}_{\mathcal{B}}$ and $\bar{\boldsymbol{H}}_{ij} \coloneqq (\bar{\boldsymbol{V}}_{i,\mathsf{e}}^{-1} + \bar{\boldsymbol{V}}_{j,\mathsf{e}}^{-1})^{-1}$,

$$\boldsymbol{U}^\top \bar{\boldsymbol{V}}_{i,\mathsf{e}}^{-1} \boldsymbol{U} = \operatorname{diag}\left(\frac{1}{\lambda + n_{i,\mathsf{e}} \zeta_b} \right)_{b \in [B]},$$
$$\boldsymbol{U}^\top \bar{\boldsymbol{H}}_{ij} \boldsymbol{U} = \operatorname{diag}\left(\frac{(\lambda + n_{i,\mathsf{e}} \zeta_b)(\lambda + n_{j,\mathsf{e}} \zeta_b)}{2\lambda + (n_{i,\mathsf{e}} + n_{j,\mathsf{e}}) \zeta_b} \right)_{b \in [B]},$$

where the second identity uses $(1/a + 1/b)^{-1} = ab/(a+b)$ for each diagonal entry. Consequently, if the implementation stores the exploration reward vector $\bar{\boldsymbol{b}}_{i,\mathsf{e}}$ of each user in this basis, the b-th coordinate of the exploration estimate of user i is obtained by a single division,

$$\widehat{\vartheta}_{i,b} = \frac{[\boldsymbol{U}^\top \bar{\boldsymbol{b}}_{i,\mathsf{e}}]_b}{\lambda + n_{i,\mathsf{e}} \zeta_b},$$

so the estimates of all users take $\mathcal{O}(NB)$ operations, and the squared harmonic-mean statistic in Eq. (4) for a pair (i, j) becomes the sum

$$\sum_{b=1}^{B} \frac{(\lambda + n_{i,\mathsf{e}} \zeta_b)(\lambda + n_{j,\mathsf{e}} \zeta_b)}{2\lambda + (n_{i,\mathsf{e}} + n_{j,\mathsf{e}}) \zeta_b} \left(\widehat{\vartheta}_{i,b} - \widehat{\vartheta}_{j,b} \right)^2.$$

Each pair is therefore tested in $\mathcal{O}(B)$ operations, and a graph update tests all remaining edges in $\mathcal{O}(N^2 B)$ operations.

Deleting an edge (i, j) removes user j from $\mathcal{N}(i)$ and user i from $\mathcal{N}(j)$, so the past observations of user j must be removed from the pooled values of user i, and vice versa. Because the per-user values are stored, this removal is a subtraction: for every arm $k \in [K]$, the implementation sets $n_{\mathcal{N}(i),k} \leftarrow n_{\mathcal{N}(i),k} - n_{j,k}$ and $Y_{\mathcal{N}(i),k} \leftarrow Y_{\mathcal{N}(i),k} - Y_{j,k}$, and symmetrically for user j, which takes $\mathcal{O}(K)$ operations and requires no past observation to be revisited. Since edges are never added back, at most $N(N-1)/2$ deletions occur over the entire horizon, so all such corrections take $\mathcal{O}(N^2 K)$ operations in total. After the deletions, the implementation rebuilds the neighbor list of every user by scanning the corresponding row of the $N \times N$ adjacency matrix, which takes $\mathcal{O}(N^2)$ operations per update; the lists allow scalar propagation to visit only the users in $\mathcal{N}(u_t)$. Graph updates occur only at phase transitions, hence at most J_T times (Lemma 8).

Total cost. The total cost of the implementation over T rounds consists of three parts. In every round, scalar propagation takes $\mathcal{O}(N)$ operations, and an adaptive round additionally takes $\mathcal{O}(KB^2)$ operations to construct and factorize the pooled design matrix of the served user. At each graph update, testing all remaining edges and rebuilding the neighbor lists take $\mathcal{O}(N^2 B)$ operations, and graph updates occur at most $J_T = \mathcal{O}(\log(T+1))$ times. Finally, each deleted edge requires a correction of $\mathcal{O}(K)$ operations, and at most $N(N-1)/2$ edges are deleted over the entire horizon. Excluding the $\mathcal{O}(NK)$ initialization and the one-time preprocessing, which selects $\mathcal{B}$, computes $\boldsymbol{Q}$ and $\boldsymbol{Z}$, and computes the eigendecomposition of $\bar{\boldsymbol{\Sigma}}_{\mathcal{B}}$ in $\mathcal{O}(KdB + B^3)$ operations, the total cost is therefore

$$\mathcal{O}\left(T(N + KB^2) + N^2 K + N^2 B \log(T+1) \right),$$

which is at most $\mathcal{O}\left(T(N + Kd^2) + N^2 K + N^2 d \log(T+1) \right)$ since $B \le d$. The number of users N enters the per-round cost only linearly, through scalar propagation.

The separation test of `BASS` (Cherkaoui et al., 2025) decides whether the confidence ellipsoids of two users intersect. Because the two ellipsoids have different shapes, the test requires matrix decompositions of the two design matrices and a one-dimensional numerical minimization; with τ iterations of the minimization, each test costs $\mathcal{O}(d^3 + \tau d)$ operations. Under the one-user-per-round protocol, `BASS` applies the test in every round to each of the up to $N-1$ edges incident to the served user, so the total cost of `BASS` is $\mathcal{O}\left(T(Kd^2 + Nd^3 + N\tau d) \right)$. `TRIBE` instead tests edges only at the $J_T = \mathcal{O}(\log(T+1))$ graph updates and tests each pair in $\mathcal{O}(B)$ operations without numerical minimization. As a result, the total cost of the separation tests decreases from $\mathcal{O}(TN(d^3 + \tau d))$ to $\mathcal{O}(N^2 d \log(T+1))$, and the dominant per-round cost decreases from $\mathcal{O}(N(d^3 + \tau d))$ to $\mathcal{O}(N + Kd^2)$.